Maximum likelihood estimation and inference for high dimensional nonlinear factor models with application to factor-augmented regressions

Fa Wang

Cass Business School, City, University of London

18 May 2017

Maximum Likelihood Estimation and Inference for High Dimensional Nonlinear Factor Models with Application to Factor-augmented Regressions*

Fa Wang†

May 19, 2019

Abstract

This paper reestablishes the main results in Bai (2003) and Bai and Ng (2006) for nonlinear factor models, with slightly stronger conditions on the relative magnitude of N(number of subjects) and T(number of time periods). Convergence rates of the estimated factor space and loading space and asymptotic normality of the estimated factors and loadings are established under mild conditions that allow for linear models, Logit, Probit, Tobit, Poisson and some other nonlinear models. The probability density/mass function is allowed to vary across subjects and time, thus mixed models are also allowed for. For factor-augmented regressions, this paper establishes the limit distributions of the parameter estimates, the conditional mean, and the forecast when factors estimated from nonlinear/mixed data are used as proxies for the true factors.

Keywords: Factor model, Mixed measurement, Maximum likelihood, High dimension, Factor-augmented regression, Forecasting

JEL Classification: C13, C35

*We would like to thank the editor Jianqing Fan, the associate editor and two referees for their valuable comments and suggestions. We would also like to thank Jushan Bai, Mingli Chen, Xiaohong Chen, Ivan Fernandez-Val, Kunpeng Li, Giovanni Urga, Martin Weidner and seminar participants at Chinese University of Hong Kong, Cass Business School, University of Connecticut and 2017 Asian Meeting of Econometric Society for helpful comments and suggestions.

†Faculty of Finance, Cass Business School, 106 Bunhill Row, London EC1Y 8TZ, U.K.

1 Introduction

High dimensional factor models where a large number of time series are simultaneously driven by a small number of latent factors provide a powerful framework to analyze high dimensional data. Accompanied by an ever-increasing data size, the literature for this model recently experienced a wave of development. For example, Bai and Ng (2002) and Bai (2003) respectively show that utilizing the high dimensionality, we are able to consistently determine the number of factors and establish the asymptotic normality of the estimated factors and loadings. High dimensional factor models have also been successfully used in macroeconomic monitoring and forecasting, business cycle analysis, asset pricing, risk measurement, see for example Stock and Watson (2002, 2016), Bernanke, Boivin and Eliasz (2005), Ross (1976) and Campbell, Lo and Mackinlay (1997), to name a few.

So far the literature only considers linear factor models. However, in many macroeconomic or financial applications and in most microeconomic applications, the relationship between the dependent variable and the factors could be nonlinear. Representative examples include but not limited to the case where the dependent variable is categorical. Direct extension of existing theory, e.g., Bai (2003) and Bai and Li (2012, 2016), to categorical data is not feasible because essentially both methods are based on the covariance matrix of the continuously distributed dependent variable. This paper seeks to establish a new estimation and inferential theory for high dimensional nonlinear factor models. More specifically, this paper considers the following single-index factor model: For $i = 1, ..., N$ and $t = 1, ..., T$,

$$x_{it} \sim g_{it}(\cdot \left| \pi_{it}^0 \right.). \tag{1}$$

x_{it} is the observed data for the i-th subject at time t. $g_{it}(\cdot \left| \cdot \right.)$ is some known probability (density or mass) function of x_{it} allowed to vary across i and t. Note that $g_{it}(\cdot \left| \cdot \right.)$ is the conditional probability function. Weak cross-sectional and serial dependence of x_{it} is allowed. $\pi_{it}^0 = f_t^{0\prime}\lambda_i^0$, and f_t^0 and λ_i^0 is an r dimensional vector of factors and loadings respectively. Both factors and loadings are unobservable. Both N and T are large. The number of factors r is known. How to determine the number of factors is studied in a separate paper.

For engineering, this model has been successfully used in data compression, visualization, pattern recognition and machine learning. For social sciences, this model also plays important role in psychology and education. For economics and finance, possible applications are partially listed below:

(1) Macroeconomic forecasting, factor-augmented vector autoregression and business cycle analysis: In these areas, common factors are predominantly estimated by principal components using continuous data, see Stock and Watson (2002), Bernanke, Boivin and Eliasz (2005) and Bai and Ng (2006). Little attention has been paid to the treatment of categorical or mixed measurement data even though many data sets are of this type. For example, let x_{1t} be the GDP, x_{2t} be the consumer confidence index (categorical), x_{3t} be the interest rate announcement of FOMC, etc, at time t. Let f_t^0 denote some macroeconomic factors, then x_{it} is nonlinearly linked to $\pi_{it}^0 = f_t^{0\prime}\lambda_i^0$ through some known link function. While mixed measurement data are quite informative, they cannot be directly handled by principal components estimation. This paper provides a rigorous solution to this issue.

(2) Credit risk analysis: Default correlation modelling has direct implications for CDO (collateralized debt obligations) pricing, bond portfolio management and commercial bank risk management. Intuitively, default correlation originates from common exposures to business cycle, monetary policy, market sentiment and other financial or sector factors. Factor models provide a parsimonious way for analyzing default correlation and underlies many risk models used in practice. In a representative case, $\pi_{it}^0 + e_{it}$ is the value of company i at time t, e_{it} is the idiosyncratic error term, f_t^0 is the common factors and x_{it} is nonlinearly linked to π_{it}^0. x_{it} could be rating category company i belongs to, or the binary variable describing the default event, or the credit spread of its bond, or its stock return, or its stock volatility at time t. For more details on default correlation modelling and estimation, see Schonbucher (2000), McNeil and Wendin (2007), Koopman and Lucas (2008), Koopman, Lucas and Monteiro (2008), Koopman, Lucas and Schwaab (2011), Creal, Schwaab, Koopman and Lucas (2014) and the references therein.

(3) Socio-economic status measurement: In development economics, health eco-

nomics, welfare economics and economics of education, researchers frequently encounter the problem of measuring the socio-economic status (more specifically the wealth or consumption) of a household or an individual. A good measure, serving as either the explanatory or the dependent variable, is crucial for these studies. Direct accurate measures of household wealth or consumption usually are not available or not reliable. Instead, the survey data contains many reliable yet categorically distributed proxies, such as living conditions and ownership of durables or assets. Treating these proxies as the dependent variables and household wealth as the latent explanatory factor, household wealth could be estimated from the data of these proxies. For example, let x_{it} be the i-th proxy of household t and let f_t^0 be the wealth of household t, then x_{it} is nonlinearly linked to $\pi_{it}^0 = f_t^{0\prime}\lambda_i^0$ through some known link function implied by economic theory. Filmer and Pritchett (2001) follows this approach to construct wealth index for estimating the effect of wealth on educational enrollments in India. The Filmer-Pritchett procedure simply extracts the factor from the binary proxies directly by principal component. Rigorously speaking, this procedure is lack of theoretical support and may lead to misleading results.

For all the above and future applications, it is in urgent need to develop a theoretically justified method for estimating the factors and loadings from high dimensional nonlinear/mixed data. It is also necessary to establish the asymptotic properties of the proposed estimator under the high dimensional setup. Such asymptotic properties are needed to characterize the conditions under which the estimation error is negligible when estimated factors are used as regressors and to construct confidence intervals when estimated factors represent economic indices.

This paper considers maximum likelihood for estimating the factors and loadings from nonlinear/mixed data. Both factors and loadings are treated as parameters to be estimated and a penalty function is added to the log-likelihood function to guarantee the uniqueness of the solution of the likelihood maximization problem. This paper establishes the convergence rates of the estimated factor space and loading space, and asymptotic normality of the estimated factors and loadings, given that the probability function satisfies some regularity conditions. These regularity conditions allow for linear models, Logit, Probit, Tobit, Poisson and some other nonlinear models. Thus

Bai (2003) is a special case of this paper. The probability function is also allowed to vary across i and t, thus a mixture of these models is allowed for. This paper also establishes the limit distributions of the parameter estimates, the conditional mean as well as the forecast for factor-augmented regression models when the estimated factors are used as proxies for the true factors. This result generalizes Bai and Ng (2006) to allow us using factors extracted from nonlinear/mixed data.

In the statistics literature, classic factor analysis has been successfully extended to categorical data and mixed data, see for example, Bartholomew (1980), Moustaki (1996), Bartholomew and Knott (1999), Moustaki (2000), Moustaki and Knott (2000) and Joreskog and Moustaki (2001), to name a few. All these papers assume N is fixed and much smaller than T. While factors are typically of primary interest in economic applications, factors can not be consistently estimated under the fixed N large T setup. This limitation and the urgent need to handle high dimensional mixed data recently has motivated researchers to explore possible solution. Ng (2015) reviews alternative methods of constructing factors that can potentially be extended to categorical data and explores their numerical properties.

This paper provides a general theory for factor analysis of high dimensional nonlinear data. Since factors and loadings are treated as parameters to be estimated, the number of parameters tend to infinity as N and T tends to infinity jointly. This paper solves this problem by utilizing the fact that for factor model, the Hessian is asymptotically block diagonal and the tensor of third order derivatives is sparse. More specifically, elements in the diagonal blocks of the Hessian are $O_p(N)$ or $O_p(T)$ while elements in the off-diagonal blocks are $O_p(1)$. This paper shows that under relevant regularity conditions, the presence of these nonzero off-diagonal blocks has no effect on the asymptotic properties of the estimated factors and loadings. Asymptotic block diagonality of the Hessian also provides explanation for Bai (2003)'s results from the perspective of extremum estimation.

This paper's solution is reminiscent of the diagonalization approaches discussed in Cox and Reid (1987) and Lancaster (2000, 2002). The difference is that in this paper the diagonality comes from the factor structure and high dimensionality and holds only when N and T tend to infinity jointly, while in those papers the diagonality comes from artificial reparametrization. More recently, Fernandez-Val and Weidner (2016) and Chen, Fernandez-Val and Weidner (2014, 2018) utilize asymptotic diagonality of the incidental parameter Hessian to derive the limit distributions of the regression coefficients and the average partial effects in nonlinear panel models. For the estimated factors and loadings, Chen et al. (2014, 2018) establishes the average consistency, while this paper also establishes the convergence rates, the limit distributions and the effect of using estimated factors in factor-augmented regression.

The rest of the paper is organized as follows. Section 2 introduces notations and preliminaries. Section 3 discusses the assumptions. Section 4 presents the limit theory. Section 5 presents results for factor-augmented regressions. Section 6 introduces computation algorithms. Section 7 presents simulation results. Section 8 concludes. All proofs are relegated to the appendix.

2 Notations and Preliminaries

The log-likelihood[1] function is

$$L(X\,|f,\lambda) = \sum\nolimits_{i=1}^{N} \sum\nolimits_{t=1}^{T} l_{it}(f_t'\lambda_i), \tag{2}$$

where $l_{it}(\pi_{it}) = \log g_{it}(x_{it}\,|\pi_{it})$ and $\pi_{it} = f_t'\lambda_i$, X is the $T \times N$ matrix of observed data and x_{it} is the element on the t-th row and the i-th column, $f = (f_1', ..., f_T')'$ a Tr dimensional vector and $\lambda = (\lambda_1', ..., \lambda_N')'$ is a Nr dimensional vector. $g_{it}(\cdot\,|\cdot)$ is allowed to vary across i and t, thus data following different models (e.g., discretely and continuously distributed time series) can be merged directly to extract common factors. We consider the following representative examples.

Example 1 *(Linear):* $l_{it}(f_t'\lambda_i) = -\frac{1}{2}(x_{it} - f_t'\lambda_i)^2$.

Example 2 *(Probit):* $l_{it}(f_t'\lambda_i) = x_{it}\log\Phi(f_t'\lambda_i) + (1-x_{it})\log(1-\Phi(f_t'\lambda_i))$, *where* $\Phi(\cdot)$ *is the CDF of the standard normal distribution.*

[1]When x_{it} is cross sectionally or serially dependent, $L(X\,|f,\lambda)$ is the quasi-likelihood function.

Example 3 *(Logit): $l_{it}(f_t'\lambda_i) = x_{it}\log\Psi(f_t'\lambda_i) + (1-x_{it})\log(1-\Psi(f_t'\lambda_i))$, where $\Psi(\cdot)$ is the CDF of the logistic distribution.*

Example 4 *(Tobit): Suppose $x_{it} = x_{it}^*$ if $x_{it}^* > 0$ and $x_{it} = 0$ if $x_{it}^* \leq 0$, where $x_{it}^* = f_t'\lambda_i + e_{it}$ and e_{it} is $N(0,1)$. The likelihood function is $l_{it}(f_t'\lambda_i) = -\frac{1}{2}(x_{it} - f_t'\lambda_i)^2\mathbf{1}(x_{it} > 0) + \log(1-\Phi(f_t'\lambda_i))\mathbf{1}(x_{it} = 0)$, where $\mathbf{1}(\cdot)$ is the indicator function.*

Example 5 *(Poisson): $l_{it}(f_t'\lambda_i) = -e^{f_t'\lambda_i} + kf_t'\lambda_i - \log k!$, because $P(x_{it} = k) = p(k,\lambda) = e^{-\lambda}\lambda^k/k!$ and $\lambda = e^{f_t'\lambda_i}$.*

Let $\phi = (\lambda', f')'$, $F = (f_1, ..., f_T)'$, $\Lambda = (\lambda_1, ..., \lambda_N)'$. Similarly, for the true values of the factors and the loadings, let $f^0 = (f_1^{0\prime}, ..., f_T^{0\prime})'$, $\lambda^0 = (\lambda_1^{0\prime}, ..., \lambda_N^{0\prime})'$, $\phi^0 = (\lambda^{0\prime}, f^{0\prime})'$, $F^0 = (f_1^0, ..., f_T^0)'$ and $\Lambda^0 = (\lambda_1^0, ..., \lambda_N^0)'$. Also, let $\partial_\pi l_{it}(\pi_{it})$, $\partial_{\pi^2} l_{it}(\pi_{it})$ and $\partial_{\pi^3} l_{it}(\pi_{it})$ be the first, second and third order derivative of $l_{it}(\cdot)$ evaluated at π_{it}, respectively. When these derivatives are evaluated at π_{it}^0, we suppress the argument and simply write $\partial_\pi l_{it}$, $\partial_{\pi^2} l_{it}$ and $\partial_{\pi^3} l_{it}$.

Both factors and loadings are treated as parameters. Note that for any F, Λ and any $r \times r$ invertible matrix G, FG and $\Lambda(G')^{-1}$ has the same likelihood as F and Λ. To uniquely fix F and Λ, we impose the normalization such that (1) $F'F$ is diagonal, (2) $\Lambda'\Lambda$ is diagonal, (3) $\frac{1}{T}F'F = \frac{1}{N}\Lambda'\Lambda$, i.e., the estimated factors and loadings are the solution of maximizing $L(X\,|f,\lambda)$ under constraints (1)-(3). As explained in Remark 1 below, the solution of this constraint maximization problem is always the same as the solution of maximizing $Q(f,\lambda) = L(X\,|f,\lambda) + P(f,\lambda)$, where

$$\begin{aligned} P(f,\lambda) \quad = \quad & -\frac{c}{8}NT\sum\nolimits_{p=1}^{r}(\frac{1}{N}\sum\nolimits_{i=1}^{N}\lambda_{ip}^2 - \frac{1}{T}\sum\nolimits_{t=1}^{T}f_{tp}^2)^2 \\ & -\frac{c}{2}\frac{T}{N}\sum\nolimits_{p=1}^{r}\sum\nolimits_{q=p+1}^{r}(\sum\nolimits_{i=1}^{N}\lambda_{ip}\lambda_{iq})^2 \\ & -\frac{c}{2}\frac{N}{T}\sum\nolimits_{p=1}^{r}\sum\nolimits_{q=p+1}^{r}(\sum\nolimits_{t=1}^{T}f_{tp}f_{tq})^2, \end{aligned} \tag{3}$$

is a penalty function, $0 < c < b_L$ and b_L is lower bound of $|\partial_{\pi^2} l_{it}(\pi_{it})|$ as presented in Assumption 2(ii) below. Thus we can consider the estimated factors and loadings as the solution of maximizing $Q(f,\lambda)$ in asymptotic analysis. For numerical computation, the algorithms in Section 6 still solves the constraint maximization problem. The normalization (1)-(3) is slightly different from the classical normalization $\frac{1}{T}F'F = I_r$ and $\Lambda'\Lambda$ being diagonal. We choose this normalization because with penalty (3), the Hessian matrix of $Q(f, \lambda)$ has some convenient structure for analyzing its asymptotic behavior. If we choose another normalization, all results of this paper still hold, except for a different rotation matrix[2].

Let $B(\mathcal{D})$ denote the neighborhood $\|f\|_\infty \leq \mathcal{D}$ and $\|\lambda\|_\infty \leq \mathcal{D}$ for some large $\mathcal{D} > 0$, and let $\hat{f} = (\hat{f}_1', ..., \hat{f}_T')'$ and $\hat{\lambda} = (\hat{\lambda}_1', ..., \hat{\lambda}_N')'$ be the solution of maximizing $Q(f, \lambda)$ within $B(\mathcal{D})$. We will explain why taking $\hat{f}$ and $\hat{\lambda}$ within $B(\mathcal{D})$ in Remark 2 below. Let $\hat{\pi}_{it} = \hat{f}_t'\hat{\lambda}_i$, $\hat{\phi} = (\hat{\lambda}', \hat{f}')'$, $\hat{F} = (\hat{f}_1, ..., \hat{f}_T)'$ and $\hat{\Lambda} = (\hat{\lambda}_1, ..., \hat{\lambda}_N)'$. The r columns of $\hat{F}$ are ordered according to their Euclidean norm, from the largest to the smallest. The r columns of $\hat{\Lambda}$ are ordered in the same way.

Throughout the paper, let $(N, T) \to \infty$ denote N and T going to infinity jointly, $\delta_{NT} = \min\{N^{\frac{1}{2}}, T^{\frac{1}{2}}\}$, $D_{NT} = \begin{bmatrix} N \times I_{Nr} & 0 \\ 0 & T \times I_{Tr} \end{bmatrix}$, $D_{TN} = \begin{bmatrix} T \times I_{Nr} & 0 \\ 0 & N \times I_{Tr} \end{bmatrix}$. $\xrightarrow{d}$ denotes convergence in distribution. "w.p.a.1" denotes "with probability approaching 1". For matrix A, let $\rho_{\min}(A)$ denote its smallest eigenvalue and $\|A\|$, $\|A\|_F$, $\|A\|_1$, $\|A\|_\infty$ and $\|A\|_{\max}$ denote its spectral norm, Frobenius norm, 1-norm, infinity norm and max norm respectively. When A has Nr rows, divide A into N blocks with each block containing r rows and let $[A]_{iq}$ denote the q-th row in the i-th block and $[A]_i = ([A]_{i1}', ..., [A]_{ir}')'$ denote the i-th block.

Remark 1 *First note that for any F and Λ, there exists a unique matrix G such that $P(FG, \Lambda(G')^{-1}) = 0$, and $P(FG, \Lambda(G')^{-1}) < 0$ for other G. If F and Λ maximizes $Q(F, \Lambda)$, then $P(F, \Lambda) = 0$ because otherwise $P(F, \Lambda) < 0$ and we can find the appropriate G such that $L(X\,|F, \Lambda) = L(X\,|FG, \Lambda(G')^{-1})$ and $P(FG, \Lambda(G')^{-1}) = 0$, which implies $Q(F, \Lambda) < Q(FG, \Lambda(G')^{-1})$, a contradiction. Thus the solution of maximizing $Q(F, \Lambda)$ is the same as the solution of maximizing $Q(F, \Lambda)$ under the constraints $P(F, \Lambda) = 0$. The latter is the same as the solution of maximizing $L(X\,|F, \Lambda)$ under the constraints $P(F, \Lambda) = 0$, which is the same as the solution of maximizing $L(X\,|F, \Lambda)$ under the constraints (1)-(3).*

[2] To show this, we first prove the results for this normalization, and then prove the results still hold after changing the rotation.

3 Assumptions

Assumption 1 *(i)* $T^{-1}F^{0\prime}F^{0} \xrightarrow{p} \Sigma_F$ *for some positive definite* Σ_F*. There exists* $M>0$ *such that* $\|f_t^0\| \leq M$ *for all* t*.*

(ii) $N^{-1}\Lambda^{0\prime}\Lambda^{0} \xrightarrow{p} \Sigma_\Lambda$ *for some positive definite* Σ_Λ*. There exists* $M>0$ *such that* $\|\lambda_i^0\| \leq M$ *for all* i*.*

Assumption 2 *(i)* $l_{it}(\cdot)$ *is three times differentiable.*

(ii) There exists $b_U > b_L > 0$ *such that* $b_L \leq -\partial_{\pi^2} l_{it}(\pi_{it}) \leq b_U$ *within a compact space of* π_{it}*.*

(iii) $|\partial_{\pi^3} l_{it}(\pi_{it})| \leq b_U$ *within a compact space of* π_{it}*.*

Assumption 3 *There exists* $M>0$ *such that for all* N *and* T*:*

(i) $\mathbb{E}(|\partial_\pi l_{it}|^\xi) \leq M$ *for some* $\xi > 14$ *and all* i *and* t*.*

(ii) $T^{-1}\sum_{s=1}^T\sum_{t=1}^T(\gamma_N(s,t))^2 \leq M$*, where* $\gamma_N(s,t) = N^{-1}\sum_{i=1}^N \mathbb{E}(\partial_\pi l_{is}\partial_\pi l_{it})$*.*

(iii) For every (t,s)*,* $\mathbb{E}(N^{-\frac{1}{2}}\sum_{i=1}^N[\partial_\pi l_{is}\partial_\pi l_{it} - \mathbb{E}(\partial_\pi l_{is}\partial_\pi l_{it})])^2 \leq M$*.*

Assumption 4 *There exists* $M>0$ *such that for some* $\zeta > 2$ *and for all* N *and* T*,*

$\mathbb{E}(N^{-1}\sum_{i=1}^N \left\|T^{-\frac{1}{2}}\sum_{t=1}^T \partial_\pi l_{it} f_t^0\right\|^\zeta) \leq M$*,*

$\mathbb{E}(T^{-1}\sum_{t=1}^T \left\|N^{-\frac{1}{2}}\sum_{i=1}^N \partial_\pi l_{it} \lambda_i^0\right\|^\zeta) \leq M$*.*

Assumption 5 *(i)* $\mathbb{E}\left\|N^{-\frac{1}{2}}T^{-\frac{1}{2}}\sum_{i=1}^N\sum_{t=1}^T(T^{-1}\sum_{t=1}^T \partial_{\pi^2}l_{it}f_t^0 f_t^{0\prime})^{-1} f_t^0 \partial_\pi l_{it}\partial_\pi l_{is}\right\|^2 \leq M$ *for any* s *and*

$\mathbb{E}\left\|N^{-\frac{1}{2}}T^{-\frac{1}{2}}\sum_{i=1}^N\sum_{t=1}^T(N^{-1}\sum_{i=1}^N \partial_{\pi^2}l_{it}\lambda_i^0 \lambda_i^{0\prime})^{-1} \lambda_i^0 \partial_\pi l_{it}\partial_\pi l_{jt}\right\|^2 \leq M$ *for any* j*.*

(ii) $\mathbb{E}\left\|N^{-\frac{1}{2}}T^{-\frac{1}{2}}\sum_{i=1}^N\sum_{t=1}^T(T^{-1}\sum_{t=1}^T \partial_{\pi^2}l_{it}f_t^0 f_t^{0\prime})^{-1} \partial_\pi l_{it} f_t^0 \lambda_i^{0\prime}\right\|^2 \leq M$ *and*

$\mathbb{E}\left\|N^{-\frac{1}{2}}T^{-\frac{1}{2}}\sum_{i=1}^N\sum_{t=1}^T(N^{-1}\sum_{i=1}^N \partial_{\pi^2}l_{it}\lambda_i^0 \lambda_i^{0\prime})^{-1} \partial_\pi l_{it} \lambda_i^0 f_t^{0\prime}\right\|^2 \leq M$*.*

$\mathbb{E}\left\|N^{-\frac{1}{2}}T^{-\frac{1}{2}}\sum_{i=1}^N\sum_{t=1}^T(T^{-1}\sum_{t=1}^T \partial_{\pi^2}l_{it}f_t^0 f_t^{0\prime})^{-1} \partial_\pi l_{it} f_t^0 \lambda_i^{0\prime}\partial_{\pi^2}l_{is}\right\|^2 \leq M$ *for any* s *and*

$\mathbb{E}\left\|N^{-\frac{1}{2}}T^{-\frac{1}{2}}\sum_{i=1}^N\sum_{t=1}^T(N^{-1}\sum_{i=1}^N \partial_{\pi^2}l_{it}\lambda_i^0 \lambda_i^{0\prime})^{-1} \partial_\pi l_{it} \lambda_i^0 f_t^{0\prime}\partial_{\pi^2}l_{jt}\right\|^2 \leq M$ *for any* j*.*

(iii) for any i*,* $-T^{-1}\sum_{t=1}^T \partial_{\pi^2}l_{it}f_t^0 f_t^{0\prime} \xrightarrow{p} \Sigma_{iF}$ *and* $T^{-\frac{1}{2}}\sum_{t=1}^T \partial_\pi l_{it} f_t^0 \xrightarrow{d} \mathcal{N}(0,\Omega_{iF})$ *for some positive definite* Σ_{iF} *and* Ω_{iF}*.*

(iv) for any t, $-N^{-1}\sum_{i=1}^{N}\partial_{\pi^2}l_{it}\lambda_i^0\lambda_i^{0\prime}\to\Sigma_{t\Lambda}$ and $N^{-\frac{1}{2}}\sum_{i=1}^{N}\partial_{\pi}l_{it}\lambda_i^0\xrightarrow{d}\mathcal{N}(0,\Omega_{t\Lambda})$ for some positive definite $\Sigma_{t\Lambda}$ and $\Omega_{t\Lambda}$.

Assumption 6 *The eigenvalues of the $r\times r$ matrix $(\Sigma_F\cdot\Sigma_\Lambda)$ are different.*

Assumption 7 $\frac{N^{\frac{3}{\xi}}T^{\frac{3}{\xi}}(N+T)^{\frac{1}{\zeta}}}{\delta_{NT}}\to 0$ *as* $(N,T)\to\infty$.

Assumption 1(i) corresponds to Assumption A in Bai (2003). Factors are allowed to be dynamic with arbitrary dynamics. Assumption 1(ii) is exactly the same as Assumption B in Bai (2003), and ensures each factor has a nontrivial contribution. Note that here $\|f_t^0\|$ and $\|\lambda_i^0\|$ are assumed to be uniformly bounded. This assumption is the same as Bai and Li (2016), but stronger than Bai (2003), which only assumes uniform boundedness of $\mathbb{E}\|f_t^0\|^4$ and $\mathbb{E}\|\lambda_i^0\|^4$. In general, compactness of parameter space is quite common for nonlinear models, e.g., Newey and McFadden (1994), Jennrich (1969) and Wu (1981). Under the current setup, this assumption is necessary because the convergence rate (and hence limit distribution) of $\hat{f}_t$ is not uniform over the parameter space of f_t^0 if $\left|\partial_{\pi^2}l_{it}(f_t^{0\prime}\lambda_i^0)\right|\to 0$ as $\|f_t^0\|\to\infty$. In other words, in such cases the convergence rates of $\hat{f}_t$ will not be the same[3] for all t.

Assumption 2(i) imposes smoothness condition on the log-likelihood function. Assumption 2(ii) and (iii) assumes that the log-likelihood function is concave, the second order derivatives are bounded below and above, and the third order derivatives are bounded above. The boundedness of the second and third order derivatives is needed to control the remainder term in the expansion of the first order condition[4]. The boundedness from below of the second order derivatives together with boundedness of π_{it} are used to show consistency of the estimated factors and loadings. We verify in Appendix D that Logit, Probit, Poisson and Tobit all satisfy Assumption 2.

[3]For example, consider the case f_t^0 is one dimensional and $\left|\partial_{\pi^2}l_{it}(f_t^{0\prime}\lambda_i^0)\right|$ converges to zero monotonically as $f_t^0\to\infty$. Let $t^*=\arg\max f_t$ and $t^{**}=\arg\min f_t$. Then convergence rate of $\hat{f}_{t^*}$ would be slower than $\hat{f}_{t^{**}}$ as $(N,T)\to\infty$.

[4]Newey and McFadden (1994) only requires two times continuously differentiable because it expands the first order condition only to the second order and utilizes Lemma 2.4 to establish the convergence of the Hessian. In this paper we expand the first order condition to the third order and utilize the uniform boundedness of the third order derivatives to explicitly calculate the magnitude the third order term. Lemma 2.4 in Newey and McFadden (1994) is no longer applicable here because the dimension of the parameter space and the dimension of the Hessian also tend to infinity.

These are most frequently used nonlinear models. For other models, readers can check accordingly.

Assumptions 3-5 are generalization of Assumptions C, D and F in Bai (2003) in the nonlinear setup. When the model is linear, $\partial_{\pi} l_{it}$ is the error term "e_{it}" and $\partial_{\pi^2} l_{it}$ is a constant, and Assumptions 3-5 reduce to Assumptions C, D and F in Bai (2003) respectively (with slight modification on the value of ξ and ζ and the statement of Assumption F1). As Bai (2003), distribution of x_{it} is allowed to be heterogeneous over i and t, and limited cross-sectional and serial dependence of x_{it} is also allowed. If x_{it} is independent over i and t conditional on the factors and loadings, Assumption 3(ii) and (iii), Assumption 4 and Assumption 5 can be easily verified. If there is no conditional independence, these assumptions still can be verified provided certain weak dependence conditions are imposed on. We follow Bai (2003)'s treatment in presenting Assumptions 3-5.

Assumption 6 is a crucial identification condition and is the same as Assumption G in Bai (2003). It guarantees that there exists unique F and Λ such that $F\Lambda' = F^0\Lambda^{0\prime}$, $F'F$ and $\Lambda'\Lambda$ are diagonl and $F'F/T = \Lambda'\Lambda/N$. Assumption 7 is quite weak if ξ and ζ are large. Note that except for some well-designed mathematical counterexamples, Assumptions 3(i) and 5 indeed hold with very large ξ and ζ.

4 Limit Theory for Estimated Factors and Loadings

For any F^0 and Λ^0, let $\rho_1^2 > ... > \rho_r^2$ be the eigenvalues of $N^{-1}T^{-1}(\Lambda^{0\prime}\Lambda^0)^{\frac{1}{2}}F^{0\prime}F^0(\Lambda^{0\prime}\Lambda^0)^{\frac{1}{2}}$ and Υ be the matrix of corresponding eigenvectors, and let $\mathcal{V} = diag(\rho_1^2, ..., \rho_r^2)$. Assumption 1 implies that $\mathcal{V}$ converges in probability to the diagonal matrix of eigenvalues of $\Sigma_\Lambda^{\frac{1}{2}}\Sigma_F\Sigma_\Lambda^{\frac{1}{2}}$ and Υ converges in probability to the matrix of eigenvectors of $\Sigma_\Lambda^{\frac{1}{2}}\Sigma_F\Sigma_\Lambda^{\frac{1}{2}}$. Let $G = (\frac{\Lambda^{0\prime}\Lambda^0}{N})^{\frac{1}{2}}\Upsilon\mathcal{V}^{-\frac{1}{4}}$, G converges in probability to a constant matrix. Assumption 6 guarantees G is unique for N and T large enough. Relationship of G and Bai (2003)'s rotation matrix will be discussed later in Proposition 5. Let

$F^G = F^0 G$ and $\Lambda^G = \Lambda^0 (G^{-1})'$. It can be easily verified that $F^G \Lambda^{G\prime} = F^0 \Lambda^{0\prime}$ and

$$\frac{1}{T} F^{G\prime} F^G = \frac{1}{N} \Lambda^{G\prime} \Lambda^G = \mathcal{V}^{\frac{1}{2}}. \tag{4}$$

Similar to the notation in Section 2, let $F^G = (f_1^G, ..., f_T^G)'$, $\Lambda^G = (\lambda_1^G, ..., \lambda_N^G)'$, $f^G = (f_1^{G\prime}, ..., f_T^{G\prime})'$, $\lambda^G = (\lambda_1^{G\prime}, ..., \lambda_N^{G\prime})'$ and $\phi^G = (\lambda^{G\prime}, f^{G\prime})'$. By definition of F^G, it is easy to see that $f_t^G = G' f_t^0$ and $\lambda_i^G = G^{-1} \lambda_i^0$. We consider ϕ^G because G is the unique rotation such that $P(\lambda^G, f^G) = 0$.

Let $S(\phi) = \partial_\phi Q(\phi)$, $S_\lambda(\phi) = \partial_\lambda Q(\phi)$ and $S_f(\phi) = \partial_f Q(\phi)$ denote the score, it follows that $S(\phi) = (S_\lambda'(\phi), S_f'(\phi))'$. Let $H(\phi) = \partial_{\phi\phi'} Q(\phi)$ be the Hessian matrix. Decomposition of $H(\phi)$ and the expression of each component is presented in Appendix A. We suppress the argument when $S(\phi)$ and $H(\phi)$ are evaluated at ϕ^G, i.e., $S = S(\phi^G)$ and $H = H(\phi^G)$.

Remark 2 *$B(\mathcal{D})$ is designed such that (1) $\hat{f}_t' \hat{\lambda}_i$ is uniformly bounded over i and t, (2) ϕ^G lies in*[5] *$B(\frac{\mathcal{D}}{2})$ w.p.a.1. Fact (1) is crucial for proving average consistency of $\hat{\phi}$, see Proposition 1 below. Fact (2) guarantees that ϕ^G lies in the interior of $B(\mathcal{D})$.*

4.1 Consistency

There are two difficulties in establishing consistency. First, the number of parameters tends to infinity jointly with N and T. Thus the classical procedure for extremum estimators, e.g., Newey and McFadden (1994), is no longer applicable. Second, the parameters are present in both dimensions and the likelihood function is nonconcave with respect to the parameters. Thus it is not feasible to extend the proof strategy of large dimensional nonlinear panels to the current setup, because they either require there is only individual effects or time effects (see for example, Hahn and Newey (2004) and Hahn and Kuersteiner (2011)), or require global concavity of the likelihood function (Fernandez-Val and Weidner (2016)). Inspired by Lemma 1 of Chen et al. (2014), this paper solves the difficulties by utilizing the boundedness from below of $-\partial_{\pi^2} l_{it}(\pi_{it})$ over the compact parameter space.

[5] Note that $\|f^G\|_\infty$ and $\left\|\lambda^G\right\|_\infty$ are bounded w.p.a.1, because f_t^0 and λ_i^0 are uniformly bounded and $\|G\|$ is bounded w.p.a.1. Thus ϕ^G lies in $B(\frac{\mathcal{D}}{2})$ w.p.a.1 when $\mathcal{D}$ is large enough.

Proposition 1 (Average Consistency) *Under Assumptions 1-3 and 6, as* $(N,T) \to \infty$, $\left\|\hat{f} - f^G\right\| = O_p(\sqrt{\frac{T}{\delta_{NT}}})$ *and* $\left\|\hat{\lambda} - \lambda^G\right\| = O_p(\sqrt{\frac{N}{\delta_{NT}}})$.

A remaining issue is that $S(\hat{\phi})$ is not necessarily zero, because the criterion function is not globally concave. If $S(\hat{\phi}) \neq 0$, then we can not utilize the first order conditions to move forward. We next show that $S(\hat{\phi}) = 0$ w.p.a.1. First, Proposition 1 implies that $\hat{\phi}$ lies in the neighborhood $\left\|D_{NT}^{-\frac{1}{2}}(\phi - \phi^G)\right\| \leq m$ w.p.a.1. By definition, $\hat{\phi}$ maximizes the likelihood within $B(\mathcal{D})$. Thus $\hat{\phi}$ maximizes the likelihood within $B(\mathcal{D}) \cap \left\|D_{NT}^{-\frac{1}{2}}(\phi - \phi^G)\right\| \leq m$ w.p.a.1. Second, we show in the Appendix that within the region $B(\mathcal{D}) \cap \left\|D_{NT}^{-\frac{1}{2}}(\phi - \phi^G)\right\| \leq m$, w.p.a.1, the criterion function is concave (see Lemma 3) and there exists a zero point of $S(\phi)$. This implies that the zero point should maximize the likelihood within $B(\mathcal{D}) \cap \left\|D_{NT}^{-\frac{1}{2}}(\phi - \phi^G)\right\| \leq m$ w.p.a.1. Thus $\hat{\phi}$ must be the zero point w.p.a.1.

Proposition 2 *Under Assumptions 1-4, 6 and 7,* $S(\hat{\phi}) = 0$ *w.p.a.1.*

All subsequent results do not rely on Assumption 7 directly. They rely on Assumption 7 purely because they rely on Proposition 2. Bai and Ng (2002) and Bai (2003) do not need any condition on the relative magnitude of N and T because in the linear setup the principal component estimator is just the global maximum, i.e., Bai and Ng (2002) and Bai (2003) do not have the difficulty[6] we encounter here.

An intermediate step for Proposition 2 is the following uniform rates.

Proposition 3 (Uniform Consistency) *Under Assumptions 1-4 and 6,*
(i) $\left\|\hat{\lambda} - \lambda^G\right\|_\infty = O_p(\frac{N^{\frac{2}{\xi}} T^{\frac{2}{\xi}} (N+T)^{\frac{1}{\zeta}}}{T^{\frac{1}{2}}})$, *(ii)* $\left\|\hat{f} - f^G\right\|_\infty = O_p(\frac{N^{\frac{3}{\xi}} T^{\frac{3}{\xi}} (N+T)^{\frac{1}{\zeta}}}{N^{\frac{1}{2}}})$.

Note that normally ξ and ζ could be large, and in such case $\left\|\hat{\lambda} - \lambda^G\right\|_\infty$ and $\left\|\hat{f} - f^G\right\|_\infty$ is approximately $O_p(T^{-\frac{1}{2}})$ and $O_p(N^{-\frac{1}{2}})$, respectively. Thus these rates are more accurate than Bai (2003)'s Proposition 2 when ξ and ζ are large.

[6] If we can find a better strategy to handle this difficulty, then we may get rid of Assumption 7.

4.2 Convergence Rates

Now we can utilize the first order conditions $S(\hat{\phi}) = 0$ to move forward. Using the integral form of the mean value theorem for vector-valued functions[7] to expand the first order conditions, we have $0 = \partial_\phi Q(\hat{\phi}) = S + \tilde{H} \times (\hat{\phi} - \phi^G)$, where $\tilde{H} = \int_0^1 H(\phi^G + s(\hat{\phi} - \phi^G))ds \equiv \int_0^1 H(s)ds$. It follows that $\hat{\phi} - \phi^G = -\tilde{H}^{-1}S$ and

$$\begin{pmatrix} N^{-\frac{1}{2}}(\hat{\lambda} - \lambda^G) \\ T^{-\frac{1}{2}}(\hat{f} - f^G) \end{pmatrix} = D_{NT}^{-\frac{1}{2}}(\hat{\phi} - \phi^G) = N^{-\frac{1}{2}}T^{-\frac{1}{2}}(-D_{TN}^{-\frac{1}{2}}\tilde{H}D_{TN}^{-\frac{1}{2}})^{-1}D_{TN}^{-\frac{1}{2}}S, \tag{5}$$

where D_{NT} and D_{TN} are normalization matrices defined in Section 2. Given Assumption 4, it is easy to see that $\left\| D_{TN}^{-\frac{1}{2}}S \right\| = O_p((N+T)^{\frac{1}{2}})$. Utilizing the structure of $H(\phi)$ and eigenvalue perturbation technique, we show in the Appendix (Lemma 3) that the largest eigenvalue of $(-D_{TN}^{-\frac{1}{2}}H(\phi)D_{TN}^{-\frac{1}{2}})^{-1}$ is $O_p(1)$ uniformly within the neighborhood $B(\mathcal{D}) \cap \left\| D_{NT}^{-\frac{1}{2}}(\phi - \phi^G) \right\| \leq m$ for some $m > 0$. Since $\hat{\phi}$ lies in $B(\mathcal{D}) \cap \left\| D_{NT}^{-\frac{1}{2}}(\phi - \phi^G) \right\| \leq m$ w.p.a.1, this implies that $\left\| (-D_{TN}^{-\frac{1}{2}}\tilde{H}D_{TN}^{-\frac{1}{2}})^{-1} \right\|$ is $O_p(1)$. Thus we have the following result:

Theorem 1 (Average Rate) *Under Assumptions 1-4, 6 and 7,* $\left\| \hat{f} - f^G \right\| = O_p(\frac{T^{\frac{1}{2}}}{\delta_{NT}})$ *and* $\left\| \hat{\lambda} - \lambda^G \right\| = O_p(\frac{N^{\frac{1}{2}}}{\delta_{NT}})$.

Theorem 1 establishes the convergence rate of the estimated factor space and the estimated loading space. In applications where estimated factors are used as proxies for the true factors, e.g., forecasting, portfolio construction, Theorem 1 provides the foundation for characterizing the effect of using estimated factors. In this paper, we shall use Theorem 1 to show the limit distributions of $\hat{\lambda}_i - \lambda_i^G$ and $\hat{f}_t - f_t^G$, and limit distribution of the parameter estimates in factor-augmented regressions.

Remark 3 *The key step for Theorem 1 is to show that* $\left\| (-D_{TN}^{-\frac{1}{2}}H(\phi)D_{TN}^{-\frac{1}{2}})^{-1} \right\|$ *is* $O_p(1)$ *uniformly within* $B(\mathcal{D}) \cap \left\| D_{NT}^{-\frac{1}{2}}(\phi - \phi^G) \right\| \leq m$. *Lemma 5 of Chen et al. (2014) proves similar result for the case of one factor. To generalize from one factor to multiple factors, there are some purely mathematical difficulties. This paper solves*

[7] Note that the standard mean value theorem does not hold for vector-valued functions. For more details, also see Feng, Wang, Han, Xia and Tu (2013).

the difficulties in step (2) of Lemma 2 and Lemma 3. Step (1) of Lemma 2 is similar to (and inspired by) Lemma 5 of Chen et al. (2014).

4.3 Limit Distributions

Now we proceed to to establish the limit distributions of the estimated factors and loadings. First, it is not feasible to extend Bai (2003)'s method of deriving the limit distribution of $\hat{f}_t - f_t^G$ to the nonlinear setup, because Bai (2003)'s method relies on expression A.1 in Appendix A of Bai (2003), a crucial decomposition identity that does not hold in nonlinear setup. Second, noting that $\hat{\lambda}_i$ can be regarded as the maximum likelihood estimator when $\hat{f}$ is used for f^G and vice versa, another choice is to expand the first order conditions $\sum_{t=1}^T \partial_\pi l_{it}(\hat{f}_t'\hat{\lambda}_i)\hat{f}_t = 0$ at λ_i^G and use Theorem 1 to study the effect of using $\hat{f}$ for f^G and $\hat{\lambda}$ for λ^G. When the model is linear, Bai (2003) uses this method to establish the limit distributions of $\hat{\lambda}_i - \lambda_i^G$. However, as explained in Remark 4 below, this method is not promising when the model is nonlinear.

Remark 4 *Using the integral form of the mean value theorem, the expansion of the first order conditions is*

$$0 = \sum_{t=1}^T \partial_\pi l_{it}(\hat{f}_t'\lambda_i^G)\hat{f}_t + \sum_{t=1}^T [\int_0^1 \partial_{\pi^2} l_{it}(\hat{f}_t'(\lambda_i^G + s(\hat{\lambda}_i - \lambda_i^G)))ds]\hat{f}_t\hat{f}_t'(\hat{\lambda}_i - \lambda_i^G). \quad (6)$$

The first term on the right hand side equals

$$\begin{aligned} &\sum_{t=1}^T (\partial_\pi l_{it}) f_t^G + \sum_{t=1}^T [\partial_\pi l_{it}(\hat{f}_t'\lambda_i^G) - \partial_\pi l_{it}] f_t^G \\ &+ \sum_{t=1}^T (\partial_\pi l_{it})(\hat{f}_t - f_t^G) + \sum_{t=1}^T [\partial_\pi l_{it}(\hat{f}_t'\lambda_i^G) - \partial_\pi l_{it}](\hat{f}_t - f_t^G). \end{aligned} \quad (7)$$

When the model is linear, without loss of generality, suppose $l_{it}(\pi_{it}) = -\frac{1}{2}(x_{it} - \pi_{it})^2$*. Then* $\partial_{\pi^2} l_{it}(\cdot)$ *always equals* -1 *and* f_t^G, λ_i^G, $\hat{f}_t$, $\hat{\lambda}_i$, $\partial_\pi l_{it}$ *and* $\partial_\pi l_{it}(\hat{f}_t'\lambda_i^G)$ *can be replaced by* "$H'F_t^0$", "$H^{-1}\lambda_i^0$", "$\tilde{F}_t$", "$\tilde{\lambda}_i$", "e_{it}" *and* "$-(\tilde{F}_t - H'F_t^0)'H^{-1}\lambda_i^0 + e_{it}$" *in Bai (2003) respectively. It follows that the four terms in expression (7) becomes* "$\sum_{t=1}^T H'F_t^0 e_{it}$", "$-\sum_{t=1}^T H'F_t^0(\tilde{F}_t - H'F_t^0)'H^{-1}\lambda_i^0$", "$\sum_{t=1}^T(\tilde{F}_t - H'F_t^0)e_{it}$" *and* "$-\sum_{t=1}^T(\tilde{F}_t - H'F_t^0)(\tilde{F}_t - H'F_t^0)'H^{-1}\lambda_i^0$" *in Bai (2003) respectively, and the second* *term on the right hand side of equation (6) becomes "$-\sum_{t=1}^{T}\tilde{F}_t\tilde{F}_t'(\tilde{\lambda}_i - H^{-1}\lambda_i^0)$" in Bai (2003). "$T^{-\frac{1}{2}}\sum_{t=1}^{T}H'F_t^0e_{it}$" is normally distributed in the limit. Lemma B.2, Lemma B.1 and Lemma A.1 in Bai (2003) shows respectively that the last three terms of expression (7) are $O_p(\frac{T}{\delta_{NT}^2})$, which is dominated by the first term if $T^{\frac{1}{2}}/N \to 0$. Lemma B.2 and Lemma A.1 in Bai (2003) also shows that "$T^{-1}\sum_{t=1}^{T}\tilde{F}_t\tilde{F}_t'$" converges in probability to some constant matrix. These together shows that "$T^{\frac{1}{2}}(\tilde{\lambda}_i - H^{-1}\lambda_i^0)$" is normally distributed in the limit.*

When the model is nonlinear, we have already reestablished Lemma A.1 of Bai (2003) in Theorem 1. It is also feasible to reestablish Lemma B.1 and Lemma B.2 of Bai (2003), as shown in Lemma 11 in the Appendix. The difficulty is that we can not get the accurate rate of the magnitude of $\sum_{t=1}^{T}[\partial_\pi l_{it}(\hat{f}_t'\lambda_i^G) - \partial_\pi l_{it}]f_t^G$, because we do not have an analytical expression for $\partial_\pi l_{it}(\hat{f}_t'\lambda_i^G) - \partial_\pi l_{it}$.

To solve this problem, we expand the first order conditions $S(\hat{\phi}) = 0$ at ϕ^G.

$$0 = S(\hat{\phi}) = S + H \times (\hat{\phi} - \phi^G) + \frac{1}{2}R,$$

where $R = (R_\lambda', R_f')'$. R_λ and R_f is Nr and Tr dimensional with element $R_{\lambda,iq} = (\hat{\phi}-\phi^G)'\partial_{\phi\phi'\lambda_{iq}}Q(\phi_{iq}^*)(\hat{\phi}-\phi^G)$ and $R_{f,tq} = (\hat{\phi}-\phi^G)'\partial_{\phi\phi' f_{tq}}Q(\phi_{tq}^*)(\hat{\phi}-\phi^G)$ respectively. ϕ_{iq}^* and ϕ_{tq}^* are linear combinations of $\hat{\phi}$ and ϕ^G. Thus

$$\begin{aligned} \hat{\phi} - \phi^G &= -H^{-1}S - \frac{1}{2}H^{-1}R, && (8)\\ \text{and } \hat{\lambda}_i - \lambda_i^G &= [\hat{\phi} - \phi^G]_i = -[H^{-1}S]_i - \frac{1}{2}[H^{-1}R]_i. && (9) \end{aligned}$$

Utilizing the structure of H, we show in Appendix C.5 that

$$[H^{-1}S]_i = (\sum_{t=1}^{T}\partial_{\pi^2}l_{it}f_t^G f_t^{G\prime})^{-1}\sum_{t=1}^{T}\partial_\pi l_{it}f_t^G + O_p(N^{-\frac{1}{2}}T^{-\frac{1}{2}}). \quad (10)$$

The intuition behind equation (10) is that H is approximately block diagonal. If the Hessian is block diagonal, asymptotic behavior of the estimates for parameters within different blocks will not affect each other. Thus as long as the dimension of each block is fixed, whether the dimension of the whole Hessian tends to infinity does not matter. In current context, H is not block diagonal, but the elements in its diagonal blocks are much larger than the elements in its off-diagonal blocks ($O_p(N^{\frac{1}{2}})$ or $O_p(T^{\frac{1}{2}})$ versus $O_p(1)$). Based on this observation and the structure of H, we show that in the expansion of $[H^{-1}S]_i$, the extra terms resulting from those nonzero off-diagonal blocks together have order $O_p(N^{-\frac{1}{2}}T^{-\frac{1}{2}})$.

Based on the structure of H, Theorem 1 and Proposition 4 presented below, we show in Appendix C.5 that

$$\left\|[H^{-1}R]_i\right\| = O_p(\frac{N^{\frac{3}{\xi}}T^{\frac{3}{\xi}}}{\delta_{NT}^2}). \tag{11}$$

Thus if $\frac{T^{\frac{1}{2}}}{\delta_{NT}^2}N^{\frac{3}{\xi}}T^{\frac{3}{\xi}} \to 0$, $\|[H^{-1}R]_i\|$ would be $o_p(T^{-\frac{1}{2}})$ and hence dominated by the first term on the right hand side of equation (10).

Proposition 4 (Individual Rate) *Under Assumptions 1-4, 6 and 7,* $\left\|\hat{\lambda}_i - \lambda_i^G\right\| = O_p(\frac{1}{\delta_{NT}})$ *for each* i *and* $\left\|\hat{f}_t - f_t^G\right\| = O_p(\frac{1}{\delta_{NT}})$ *for each* t.

Remark 5 *The proof of Proposition 4 is based on expression (7) and utilizes Cauchy-Schwarz inequality and Theorem 1. The rate* $O_p(\frac{1}{\delta_{NT}})$ *is not sharp, but enough for calculating the order of* $[H^{-1}R]_i$.

Remark 6 *The reason that the remainder term* $[H^{-1}R]_i$ *is asymptotically negligible is because the tensor of third order derivatives is sparse. For example, it's easy to see that* $\sum_{i=1}^N\sum_{t=1}^T \partial_{\lambda_k\lambda_j f_s} l_{it}(\cdot) = 0$ *if* $k \neq j$, *and* $\sum_{i=1}^N\sum_{t=1}^T \partial_{\lambda_k f_l f_s} l_{it}(\cdot) = 0$ *if* $l \neq s$.

From equations (10) and (11), and the symmetry between $\hat{\lambda}_i$ and $\hat{f}_t$, we have the the following theorem.

Theorem 2 (Individual Limit Distribution) *Under Assumptions 1-7,*

$$\begin{aligned} T^{\frac{1}{2}}(\hat{\lambda}_i - \lambda_i^G) &\xrightarrow{d} \mathcal{N}(0, \bar{G}^{-1}\Sigma_{iF}^{-1}\Omega_{iF}\Sigma_{iF}^{-1}\bar{G}'^{-1}) \text{ if } \frac{T^{\frac{1}{2}}}{\delta_{NT}^2}N^{\frac{3}{\xi}}T^{\frac{3}{\xi}} \to 0, \\ N^{\frac{1}{2}}(\hat{f}_t - f_t^G) &\xrightarrow{d} \mathcal{N}(0, \bar{G}'\Sigma_{t\Lambda}^{-1}\Omega_{t\Lambda}\Sigma_{t\Lambda}^{-1}\bar{G}) \text{ if } \frac{N^{\frac{1}{2}}}{\delta_{NT}^2}N^{\frac{3}{\xi}}T^{\frac{3}{\xi}} \to 0, \end{aligned}$$

where $\bar{G} = plimG$, and Σ_{iF}, Ω_{iF}, $\Sigma_{t\Lambda}$ and $\Omega_{t\Lambda}$ are defined in Assumption 5. Asymptotic variance of $\hat{\lambda}_i$ and $\hat{f}_t$ can be estimated by

$$\begin{aligned} var_\lambda &= T(\sum_{t=1}^{T} \partial_{\pi^2} l_{it}(\hat{f}_t'\hat{\lambda}_i)\hat{f}_t\hat{f}_t')^{-1}(\sum_{t=1}^{T}(\partial_\pi l_{it}(\hat{f}_t'\hat{\lambda}_i))^2\hat{f}_t\hat{f}_t')(\sum_{t=1}^{T} \partial_{\pi^2} l_{it}(\hat{f}_t'\hat{\lambda}_i)\hat{f}_t\hat{f}_t')^{-1}, \\ var_f &= N(\sum_{i=1}^{N} \partial_{\pi^2} l_{it}(\hat{f}_t'\hat{\lambda}_i)\hat{\lambda}_i\hat{\lambda}_i')^{-1}(\sum_{i=1}^{N}(\partial_\pi l_{it}(\hat{f}_t'\hat{\lambda}_i))^2\hat{\lambda}_i\hat{\lambda}_i')(\sum_{i=1}^{N} \partial_{\pi^2} l_{it}(\hat{f}_t'\hat{\lambda}_i)\hat{\lambda}_i\hat{\lambda}_i')^{-1}. \end{aligned}$$

Theorem 2 not only allows discrete dependent variables but also allows the probability function to differ across individuals and time. The huge amount of discrete data in macroeconomic and financial studies thus can be utilized, either by themselves or merged with continuous data, to extract information on common shocks or the state of the economy or other relevant variables. In real applications, we may simply choose normal density for continuous x_{it}. For discrete x_{it}, specific parametric model is needed.

Theorem 2 allows us to construct confidence intervals for the true factor process. This is useful since in various applications factors represent economic indices. Theorem 2 also has implication for factor-augmented forecasting. Since the estimated factors will be used as proxies for true factors, the estimation error $\hat{f}_t - f_t^G$ will be reflected in the forecasting error. We shall study this in Section 5.

Remark 7 *To have limit normal distribution, Bai (2003) assumes $T^{\frac{1}{2}}/N \to 0$ for estimated loadings and $N^{\frac{1}{2}}/T \to 0$ for estimated factors. It is not difficult to see that when ξ is large, our condition is approximately the same as Bai (2003)'s condition.*

Remark 8 *"$N^{\frac{3}{\xi}}T^{\frac{3}{\xi}}$" appears because we choose to calculate $\|R\|_1$ rather than $\|R\|$. If we choose to calculate $\|R\|$, then due to the presence of the term "L1i" in Lemma 9 in the Appendix, we need to calculate the exact rate of $\left\|\hat{\lambda} - \lambda^G\right\|_4$, which seems infeasible (Note that unlike the linear case, we do not have accurate analytical expression of $\hat{\lambda}_i - \lambda_i^G$). If the model is linear, then $\partial_{\pi^3} l_{it}(\cdot) = 0$ and "L1i" would disappear, then there is no need to calculate $\|R\|_1$ and "$N^{\frac{3}{\xi}}T^{\frac{3}{\xi}}$" in all results of this paper except for Proposition 3 would also disappear.*

Remark 9 *Let $\bar{\mathcal{V}}$ =plim$\mathcal{V}$. If the model is linear, $\bar{G}'\Sigma_{iF}\bar{G} = \bar{\mathcal{V}}^{\frac{1}{2}}$ and $\bar{G}^{-1}\Sigma_{t\Lambda}\bar{G}'^{-1} = \bar{\mathcal{V}}^{\frac{1}{2}}$, and the limit variance of $\hat{\lambda}_i - \lambda_i^G$ and $\hat{f}_t - f_t^G$ become $\bar{\mathcal{V}}^{-\frac{1}{2}}\bar{G}'\Omega_{iF}\bar{G}\bar{\mathcal{V}}^{-\frac{1}{2}}$ and*

$\bar{\mathcal{V}}^{-\frac{1}{2}}\bar{G}^{-1}\Omega_{t\Lambda}\bar{G}'^{-1}\bar{\mathcal{V}}^{-\frac{1}{2}}$ respectively. If $\Sigma_{iF} = \Omega_{iF}$ and $\Sigma_{t\Lambda} = \Omega_{t\Lambda}$, the limit variance of $\hat{\lambda}_i - \lambda_i^G$ and $\hat{f}_t - f_t^G$ becomes $\bar{G}^{-1}\Sigma_{iF}^{-1}\bar{G}'^{-1}$ and $\bar{G}'\Sigma_{t\Lambda}^{-1}\bar{G}$ respectively.

4.4 Relationship of G and Bai (2003)'s Rotation Matrix

Bai (2003)'s rotation matrix is $H_{Bai} \equiv \frac{\Lambda^{0\prime}\Lambda^0}{N}\frac{F^{0\prime}\tilde{F}}{T}\mathcal{V}_{NT}^{-1}$, where $\tilde{F} = \hat{F}\mathcal{V}_{NT}^{-\frac{1}{4}}$, $\mathcal{V}_{NT} = diag(\hat{\rho}_1^2, ..., \hat{\rho}_r^2)$ and $\hat{\rho}_1 > ... > \hat{\rho}_r$ are the singular values of $N^{-\frac{1}{2}}T^{-\frac{1}{2}}\hat{F}\hat{\Lambda}'$. G depends only on f^0 and λ^0, while H_{Bai} depends not only on f^0 and λ^0 but also on the dependent variable. Moreover, we show in Appendix C.6 that

Proposition 5 *Under Assumptions 1-4, 6 and 7,*

$$\begin{aligned} \|\mathcal{V}_{NT} - \mathcal{V}\| &= O_p(\frac{N^{\frac{3}{\xi}}T^{\frac{3}{\xi}}}{\delta_{NT}^2}) && (12) \\ \left\|G\mathcal{V}_{NT}^{-\frac{1}{4}} - H_{Bai}\right\| &= O_p(\frac{N^{\frac{3}{\xi}}T^{\frac{3}{\xi}}}{\delta_{NT}^2}). && (13) \end{aligned}$$

Theorem 1 in Bai and Ng (2002) and Lemma A.1 in Bai (2003) shows $\left\|\tilde{F} - F^0H_{Bai}\right\|$ is $O_p(\frac{T^{\frac{1}{2}}}{\delta_{NT}})$, while Theorem 1 shows $\left\|\hat{F} - F^0G\right\|$ is $O_p(\frac{T^{\frac{1}{2}}}{\delta_{NT}})$. Given expressions (12)-(13) and $\tilde{F} = \hat{F}\mathcal{V}_{NT}^{-\frac{1}{4}}$, it's easy to see that $\left\|\tilde{F} - F^0H_{Bai}\right\| \leq \left\|\hat{F} - F^0G\right\| \left\|\mathcal{V}_{NT}^{-\frac{1}{4}}\right\| + T^{\frac{1}{2}}O_p(\frac{N^{\frac{3}{\xi}}T^{\frac{3}{\xi}}}{\delta_{NT}^2})$. Under Assumption 7, $O_p(\frac{N^{\frac{3}{\xi}}T^{\frac{3}{\xi}}}{\delta_{NT}}) = o_p(1)$, thus the result of Bai and Ng is a corollary (and thus special case) of Theorem 1.

Corollary 1 *Under Assumptions 1-4, 6 and 7, $\left\|\tilde{F} - F^0H_{Bai}\right\| = O_p(\frac{T^{\frac{1}{2}}}{\delta_{NT}})$.*

Theorem 1 and Theorem 2 in Bai (2003) shows that $N^{\frac{1}{2}}(\tilde{f}_t - H'_{Bai}f_t^0)$ and $T^{\frac{1}{2}}(\tilde{\lambda}_i - H_{Bai}^{-1}\lambda_i^0)$ has limit normal distribution, while Theorem 2 shows that $N^{\frac{1}{2}}(\hat{f}_t - G'f_t^0)$ and $T^{\frac{1}{2}}(\hat{\lambda}_i - G^{-1}\lambda_i^0)$ has limit normal distribution. Since $\tilde{f}_t - H'_{Bai}f_t^0 = \mathcal{V}_{NT}^{-\frac{1}{4}}(\hat{f}_t - G'f_t^0) + (G\mathcal{V}_{NT}^{-\frac{1}{4}} - H_{Bai})'f_t^0$, expressions (12)-(13) and the condition $\frac{N^{\frac{1}{2}}}{\delta_{NT}^2}N^{\frac{3}{\xi}}T^{\frac{3}{\xi}} \to 0$ implies that Bai's result is a corollary (and thus special case) of Theorem 2.

Corollary 2 *Under Assumptions 1-7,*

$$T^{\frac{1}{2}}(\tilde{\lambda}_i - H_{Bai}^{-1}\lambda_i^0) \xrightarrow{d} \mathcal{N}(0, \bar{\mathcal{V}}^{\frac{1}{4}}\bar{G}^{-1}\Sigma_{iF}^{-1}\Omega_{iF}\Sigma_{iF}^{-1}\bar{G}'^{-1}\bar{\mathcal{V}}^{\frac{1}{4}}) \text{ if } \frac{T^{\frac{1}{2}}}{\delta_{NT}^2}N^{\frac{3}{\xi}}T^{\frac{3}{\xi}} \rightarrow 0,$$
$$N^{\frac{1}{2}}(\tilde{f}_t - H_{Bai}'f_t^0) \xrightarrow{d} \mathcal{N}(0, \bar{\mathcal{V}}^{-\frac{1}{4}}\bar{G}'\Sigma_{t\Lambda}^{-1}\Omega_{t\Lambda}\Sigma_{t\Lambda}^{-1}\bar{G}\bar{\mathcal{V}}^{-\frac{1}{4}}) \text{ if } \frac{N^{\frac{1}{2}}}{\delta_{NT}^2}N^{\frac{3}{\xi}}T^{\frac{3}{\xi}} \rightarrow 0.$$

5 Inference and Forecasting for Factor-augmented Regressions

In this section we shall use the results and techniques developed in Section 4 to study the effect of using estimated factors on factor-augmented regressions. Consider the following factor-augmented regression model:

$$y_{t+h} = \alpha' f_t^0 + \beta' W_t + \epsilon_{t+h}, \tag{14}$$

where f_t^0 is a r dimensional vector of factors, W_t is a q dimensional vector of other variables and h is the lead time between the dependent variable and information available. W_t and y_{t+h} are both observable. f_t^0 is unobservable, but a large number of predictors $x_{it}(i = 1, ..., N; t = 1, ..., T)$ are observable and can be used to estimate f_t^0. The probability function of x_{it} is $g_{it}(\cdot \,|f_t^{0\prime}\lambda_i^0)$, as introduced in Section 1. $g_{it}(\cdot\,|\cdot)$ satisfies the regularity conditions listed in Assumption 2.

When y_{t+h} is a scalar and $x_{it} = f_t^{0\prime}\lambda_i^0 + e_{it}$, this is the "diffusion index forecasting model" of Stock and Watson (2002). When $h = 1$ and $y_{t+1} = (f_{t+1}^{0\prime}, W_{t+1}')'$, this is the FAVAR of Bernanke et al. (2005). When $h = 0$, y_t is a scalar and x_{it} is discretely distributed, this is the model considered in Filmer and Pritchett (2001). When y_{t+h} is a scalar and x_{it} is discretely distributed for some i and continuously distributed for the other i, this model can be used to analyze and forecast credit risk.

We shall use $\hat{F}$ as proxy for F^0. The objective is to characterize the effect of using $\hat{F}$ for F^0 on the limit distributions of the parameter estimates, the conditional mean as well as the forecast. Bai and Ng (2006) studies this effect when the factors are estimated by principal components and $x_{it} = f_t^{0\prime}\lambda_i^0 + e_{it}$. The results in this section generalize Bai and Ng (2006)'s results to allow x_{it} to have nonlinear relationship with the factors for all or some i.

Assumption 8 *Let $z_t = (f_t^{0\prime}, W_t')'$. $\mathbb{E}\left\|W_t\right\|^{\xi} \leq M$ and $\mathbb{E}(\epsilon_t^{\xi}) \leq M$ for some $\xi > 14$ and all t. $\mathbb{E}(\epsilon_{t+h}\,|y_t, z_t, y_{t-1}, z_{t-1}, \dots) = 0$ for all $h > 0$. ϵ_t is independent with x_{is} for all i and s. Furthermore,*

(i) $T^{-1}\sum_{t=1}^{T} z_t z_t' \xrightarrow{p} \Sigma_{zz}$,

(ii) $T^{-\frac{1}{2}}\sum_{t=1}^{T} z_t\epsilon_{t+h} \xrightarrow{d} \mathcal{N}(0, \Sigma_{zz\epsilon})$, where $\Sigma_{zz\epsilon} = plim T^{-1}\sum_{t=1}^{T}\epsilon_{t+h}^2 z_t z_t'$.

(iii) $\mathbb{E}\left\|N^{-\frac{1}{2}}T^{-\frac{1}{2}}\sum_{i=1}^{N}\sum_{t=1}^{T}(N^{-1}\sum_{i=1}^{N}\partial_{\pi^2}l_{it}\lambda_i^0\lambda_i^{0\prime})^{-1}\partial_{\pi}l_{it}\lambda_i^0 W_t'\right\|^2 \leq M$,

$\mathbb{E}\left\|N^{-\frac{1}{2}}T^{-\frac{1}{2}}\sum_{i=1}^{N}\sum_{t=1}^{T}(N^{-1}\sum_{i=1}^{N}\partial_{\pi^2}l_{it}\lambda_i^0\lambda_i^{0\prime})^{-1}\partial_{\pi}l_{it}\lambda_i^0\epsilon_{t+h}\right\|^2 \leq M$.

Assumption 8 corresponds to Assumption E in Bai and Ng (2006). Part (i) and part (ii) are exactly the same as part (1) and (2) of Assumption E in Bai and Ng (2006). Bai and Ng (2006) also assumes that W_t and ϵ_t are independent with "e_{is}" for all i and s, where "e_{is}" is the error term. The independence between ϵ_t and x_{is} here corresponds to their independence between ϵ_t and "e_{is}". The second condition of Assumption 8(iii) is not difficult to verify using the independence between ϵ_t and x_{is}. The first condition of Assumption 8(iii) corresponds to the independence between W_t and "e_{is}" in Bai and Ng (2006).

We shall only consider the case where y_t is a scalar. When y_t is a vector, the results are conceptually the same. Let $\hat{z}_t = (\hat{f}_t', W_t')'$ and $\delta = ((G^{-1}\alpha)', \beta')'$. Let $\hat{\delta} = (\hat{\alpha}', \hat{\beta}')'$ be the least squares estimator of regressing y_{t+h} on $\hat{z}_t$, i.e., $\hat{\alpha}$ is an estimates of $G^{-1}\alpha$.

Theorem 3 (Inference) *Under Assumptions 1-4, 6-8, and assume $\frac{T^{\frac{1}{2}}}{\delta_{NT}^2}N^{\frac{3}{\xi}}T^{\frac{4}{\xi}} \to 0$ as $(N, T) \to \infty$,*

$$T^{\frac{1}{2}}(\hat{\delta} - \delta) \xrightarrow{d} N(0, \Sigma_{\delta}),$$

where $\Sigma_{\delta} = \bar{\Xi}^{-1}\Sigma_{zz}^{-1}\Sigma_{zz\epsilon}\Sigma_{zz}^{-1}\bar{\Xi}'^{-1}$ and $\bar{\Xi} = diag(\bar{G}, I_q)$. A consistent estimator of Σ_{δ} is $\hat{\Sigma}_{\delta} = (T^{-1}\sum_{t=1}^{T-h}\hat{z}_t\hat{z}_t')^{-1}(T^{-1}\sum_{t=1}^{T-h}\hat{\epsilon}_{t+h}^2\hat{z}_t\hat{z}_t')(T^{-1}\sum_{t=1}^{T-h}\hat{z}_t\hat{z}_t')^{-1}$.

Theorem 3 implies that using the estimated factors does not affect the limit distribution of $\hat{\delta}$ when the factors are estimated by maximum likelihood and the probability function of x_{it} satisfy Assumptions 2. Theorem 3 generalizes Theorem 1 of Bai and

Ng (2006) to allow factors to be extracted from discrete or some other nonlinear data. This generalization should be valuable as in many factor-augmented regressions the information about the common factors are contained in discrete or mixed data. Theorem 3 provides theoretical support and guidance for exploiting these information.

For factor-augmented vector autoregression (FAVAR), the result and proof is conceptually the same. We do not repeat here. Thus Theorem 2 of Bai and Ng (2006) is also a special case of this paper.

Remark 10 *Theorem 1 of Bai and Ng (2006) requires $T^{\frac{1}{2}}/N \to 0$. When ξ is large, the condition $\frac{T^{\frac{1}{2}}}{\delta_{NT}^2} N^{\frac{3}{\xi}} T^{\frac{4}{\xi}} \to 0$ are close to $T^{\frac{1}{2}}/N \to 0$.*

Now consider forecasting for factor-augmented regression models. By Assumption 8, $\mathbb{E}(\epsilon_{t+h} \,|y_t, z_t, y_{t-1}, z_{t-1}, \dots) = 0$. Thus the conditional mean $y_{T+h|T}$ equals $\alpha' f_T^0 + \beta' W_T$. Let $\hat{y}_{T+h|T} = \hat{\delta}' \hat{z}_T$ be the forecast of $y_{T+h|T}$.

Theorem 4 (Forecasting) *Under Assumptions 1-8 and assume $\frac{T^{\frac{1}{2}}}{\delta_{NT}^2} N^{\frac{3}{\xi}} T^{\frac{4}{\xi}} \to 0$ and $\frac{N^{\frac{1}{2}}}{\delta_{NT}^2} N^{\frac{3}{\xi}} T^{\frac{3}{\xi}} \to 0$ as $(N, T) \to \infty$,*

$$(\hat{y}_{T+h|T} - y_{T+h|T})/B_T \xrightarrow{d} \mathcal{N}(0,1),$$

where $B_T^2 = T^{-1} z_T' \Sigma_{zz}^{-1} \Sigma_{zz\epsilon} \Sigma_{zz}^{-1} z_T + N^{-1} \alpha' \Sigma_{t\Lambda}^{-1} \Omega_{t\Lambda} \Sigma_{t\Lambda}^{-1} \alpha$. A consistent estimator of B_T^2 is $\hat{B}_T^2 = T^{-1} \hat{z}_T' \hat{\Sigma}_\delta \hat{z}_T + N^{-1} \hat{\alpha}' var_f^{-1} \hat{\alpha}$.

Theorem 4 generalizes Theorem 3 of Bai and Ng (2006) to allow factors to be extracted from discrete or some other nonlinear data. The variance of the estimated conditional mean has two components, one from the estimated parameters $\hat{\delta}$ and the other one from the estimated factors $\hat{f}_T$. Compared to cases where factors are observable, the presence of the latter component is the effect of using estimated factors on the estimated conditional mean.

Since $y_{T+h} = y_{T+h|T} + \epsilon_{T+h}$, the forecasting error is

$$\hat{\epsilon}_{T+h} = \hat{y}_{T+h|T} - y_{T+h|T} - \epsilon_{T+h}.$$

Given Theorem 4 and assume ϵ_t is $i.i.d.\mathcal{N}(0,\sigma_\epsilon^2)$, we have $\hat{\epsilon}_{T+h} \sim \mathcal{N}(0, \sigma_\epsilon^2 + var(\hat{y}_{T+h|T}))$. σ_ϵ^2 can be consistency estimated by $T^{-1}\sum_{t=1}^{T}\hat{\epsilon}_t^2$ and $var(\hat{y}_{T+h|T})$ can be consistently estimated by $\hat{B}_T^2$. Prediction intervals can be constructed correspondingly.

Remark 11 *Theorem 3 of Bai and Ng (2006) requires $T^{\frac{1}{2}}/N \to 0$ and $N^{\frac{1}{2}}/T \to 0$. When ξ is large, the conditions $\frac{T^{\frac{1}{2}}}{\delta_{NT}^2}N^{\frac{3}{\xi}}T^{\frac{4}{\xi}} \to 0$ and $\frac{N^{\frac{1}{2}}}{\delta_{NT}^2}N^{\frac{3}{\xi}}T^{\frac{3}{\xi}} \to 0$ are close to $T^{\frac{1}{2}}/N \to 0$ and $N^{\frac{1}{2}}/T \to 0$.*

6 Algorithms

We shall introduce two algorithms, alternating maximization and minorization maximization, to numerically calculate the maximum likelihood estimator. The latter is computationally simpler, but so far we can only show it applies to Probit, Logit and Tobit. Whether it applies to more general models is unknown.

6.1 Alternating Maximization (AM)

Algorithm:

Step 1 (Initial values): Randomly generate initial values of the factors, $\hat{f}^{(0)}$.

Step 2 (Iterate): For $k = 0, ...,$ calculate

$$\begin{aligned}\hat{\lambda}^{(k)} &= \arg\max L(X \,\Big|\, \hat{f}^{(k)}, \lambda), \\ \hat{f}^{(k+1)} &= \arg\max L(X \,\Big|\, f, \hat{\lambda}^{(k)}).\end{aligned}$$

Iterate until $L(X \,\Big|\, \hat{f}^{(k+1)}, \hat{\lambda}^{(k+1)}) - L(X \,\Big|\, \hat{f}^{(k)}, \hat{\lambda}^{(k)}) \leq error$, where error is the level of tolerated numerical error.

Step 3 (Repeat): Repeat step 1 and step 2 many times to get many local maximum. Take the one with the largest likelihood.

Step 4 (Normalize): Suppose $\hat{f}^{(s)}$ and $\hat{\lambda}^{(s)}$ be the estimator from step 3. Let $\hat{F}^{(s)} = (\hat{f}_1^{(s)}, ..., \hat{f}_T^{(s)})'$ and $\hat{\Lambda}^{(s)} = (\hat{\lambda}_1^{(s)}, ..., \hat{\lambda}_N^{(s)})'$. Let $\hat{V}^{(s)}$ be the diagonal matrix of eigenvalues of $N^{-1}T^{-1}(\hat{\Lambda}^{(s)\prime}\hat{\Lambda}^{(s)})^{\frac{1}{2}}\hat{F}^{(s)\prime}\hat{F}^{(s)}(\hat{\Lambda}^{(s)\prime}\hat{\Lambda}^{(s)})^{\frac{1}{2}}$ and $\hat{\Upsilon}^{(s)}$ be the corresponding matrix of *eigenvectors, and let* $\hat{G}^{(s)} = (\frac{1}{N}\hat{\Lambda}^{(s)\prime}\hat{\Lambda}^{(s)})^{\frac{1}{2}}\hat{\Upsilon}^{(s)}(\hat{V}^{(s)})^{-\frac{1}{4}}$*. Choose* $\hat{F} = \hat{F}^{(s)}\hat{G}^{(s)}$ *and* $\hat{\Lambda} = \hat{\Lambda}^{(s)}((\hat{G}^{(s)})^{-1})'$ *as the solution of the likelihood maximization problem.*

This algorithm is not totally new. In the machine learning literature, similar algorithm has been proposed in Collins, Dasgupta and Schapire (2001) and Schein, Saul and Ungar (2003). The name "Alternating Maximization" comes from step 2, where we choose $\hat{\lambda}^{(k)}$ to maximize the likelihood for given $\hat{f}^{(k)}$ and then choose $\hat{f}^{(k+1)}$ to maximize the likelihood for given $\hat{\lambda}^{(k)}$. This is based on the fact that $L(X\,|f,\lambda)$ is globally concave with respect to λ for given f and vice versa. Because the likelihood is maximized alternately, we have $L(X\left|\hat{f}^{(k+1)},\hat{\lambda}^{(k+1)}\right.) \geq L(X\left|\hat{f}^{(k+1)},\hat{\lambda}^{(k)}\right.) \geq L(X\left|\hat{f}^{(k)},\hat{\lambda}^{(k)}\right.)$. Thus convergence of step 2 to a local maximum is guaranteed.

Whether the local maximum is global depends on the initial values $(\hat{f}^{(0)},\hat{\lambda}^{(0)})$. To search the global maximum, a common practice is to randomly choose initial values many times and take the one with the largest likelihood among all local maximum. We follow this common practice in step 3. Step 4 normalizes the estimator from step 3 so that $\hat{F}'\hat{F}$ equals $\hat{\Lambda}'\hat{\Lambda}$ and both are diagonal.

6.2 Minorization Maximization (MM)

Algorithm:

Step 1 (Initial values): Randomly generate initial values of the factors and the loadings, $(\hat{f}^{(0)},\hat{\lambda}^{(0)})$.

Step 2 (Iterate): For $k = 0, ...,$ *first calculate* $\hat{x}_{it}^{(k)} = \hat{f}_t^{(k)\prime}\hat{\lambda}_i^{(k)} + \frac{1}{b_U}\partial_\pi l_{it}(\hat{f}_t^{(k)\prime}\hat{\lambda}_i^{(k)})$ *for* $i = 1, ..., N$ *and* $t = 1, ..., T$*, then* $(\hat{f}^{(k+1)},\hat{\lambda}^{(k+1)}) = \arg\min \sum_{i=1}^N \sum_{t=1}^T (\hat{x}_{it}^{(k)} - f_t'\lambda_i)^2$*. Iterate until* $L(X\left|\hat{f}^{(k+1)},\hat{\lambda}^{(k+1)}\right.) - L(X\left|\hat{f}^{(k)},\hat{\lambda}^{(k)}\right.) \leq error$*, where error is the level of tolerated numerical error.*

Step 3 (Repeat): Repeat step 1 and step 2 many times to get many local maximum. Take the one with the largest likelihood.

Step 4 (Normalize): Suppose $\hat{f}^{(s)}$ *and* $\hat{\lambda}^{(s)}$ *be the estimator from step 3. Define* $\hat{F}^{(s)}$*,* $\hat{\Lambda}^{(s)}$ *and* $\hat{G}^{(s)}$ *in the same way as step 4 of the AM algorithm. Choose* $\hat{F} = \hat{F}^{(s)}\hat{G}^{(s)}$ *and* $\hat{\Lambda} = \hat{\Lambda}^{(s)}((\hat{G}^{(s)})^{-1})'$ *as the solution of the likelihood maximization problem.*

In the econometrics literature, Chen (2016) first proposes this algorithm for nonlinear panel models. This algorithm is also studied by de Leeuw (2006) in the statistics literature. Minorization maximization is a class of algorithm more general than the expectation maximization (EM). A function $h(x|y)$ is said to minorize a function $l(x)$ at y if $h(x|y) \leq l(x)$ for all x and $h(y|y) = l(y)$, i.e., $h(x|y)$ lies below $l(x)$ and is tangent to $l(x)$ at the point y. To maximize $l(x)$, the MM algorithm starts from an initial value $x^{(0)}$ and iteratively maximizes $h(x|x^{(k)})$ until convergence. By definition of $h(x|y)$, it is not difficult to see that $l(x^{(k)}) = h(x^{(k)}|x^{(k)}) \leq h(x^{(k+1)}|x^{(k)}) \leq l(x^{(k+1)})$. Thus convergence to local maximum is guaranteed. In applications, how to choose $h(x|y)$ mainly depends on computational simplicity. If there exists a function $w(y)$ such that $l(x) - l(y) \geq l'(y)(x-y) + \frac{1}{2}w(y)(x-y)^2$ for all x and y, a popular choice is $h(x|y) = l(y) + l'(y)(x-y) + \frac{1}{2}w(y)(x-y)^2$. For more details on the MM algorithm, see Bohning and Lindsay (1988), Hunter and Lange (2004) and Lange, Hunter and Young (2000), to name a few.

In current context, in view of the fact $\partial_{\pi^2} l_{it}(\pi_{it}) \geq -b_U$ (As shown in Appendix D, $b_U = 1$ for Probit model and $b_U = \frac{1}{4}$ for Logit model.), we choose $h_{it}(x|y) = l_{it}(y) + l'_{it}(y)(x-y) - \frac{1}{2}b_U(x-y)^2$ for each (i,t). Let $\hat{\pi}_{it}^{(k)} = \hat{f}_t^{(k)\prime}\hat{\lambda}_i^{(k)}$, it follows that

$$
\begin{aligned}
l_{it}(\hat{\pi}_{it}^{(k+1)}) &\geq l_{it}(\hat{\pi}_{it}^{(k)}) + \partial_\pi l_{it}(\hat{\pi}_{it}^{(k)})(\hat{\pi}_{it}^{(k+1)} - \hat{\pi}_{it}^{(k)}) - \frac{1}{2}b_U(\hat{\pi}_{it}^{(k+1)} - \hat{\pi}_{it}^{(k)})^2 \\
&= l_{it}(\hat{\pi}_{it}^{(k)}) - \frac{1}{2}b_U(\hat{\pi}_{it}^{(k+1)} - \hat{\pi}_{it}^{(k)} - \frac{\partial_\pi l_{it}(\hat{\pi}_{it}^{(k)})}{b_U})^2 + \frac{(\partial_\pi l_{it}(\hat{\pi}_{it}^{(k)}))^2}{2b_U}.
\end{aligned}
$$

Take sum over i and t, then $L(X \,\Big|\, \hat{f}^{(k+1)}, \hat{\lambda}^{(k+1)}) - L(X \,\Big|\, \hat{f}^{(k)}, \hat{\lambda}^{(k)})$ is not smaller than

$$
-\frac{1}{2}b_U \sum_{i=1}^{N}\sum_{t=1}^{T}(\hat{x}_{it}^{(k)} - \hat{\pi}_{it}^{(k+1)})^2 + \frac{1}{2b_U}\sum_{i=1}^{N}\sum_{t=1}^{T}(\partial_\pi l_{it}(\hat{\pi}_{it}^{(k)}))^2.
$$

If $\hat{\pi}_{it}^{(k+1)} = \hat{\pi}_{it}^{(k)}$, this term is zero. Since $\hat{f}_t^{(k+1)}$ and $\hat{\lambda}_i^{(k+1)}$ minimizes $\sum_{i=1}^{N}\sum_{t=1}^{T}(\hat{x}_{it}^{(k)} - f_t'\lambda_i)^2$, this term must be nonnegative, and consequently $L(X \,\Big|\, \hat{f}^{(k+1)}, \hat{\lambda}^{(k+1)})$ is not smaller than $L(X \,\Big|\, \hat{f}^{(k)}, \hat{\lambda}^{(k)})$. This guarantees convergence of step 2 to a local maximum. Step 3 and Step 4 are the same as the AM algorithm discussed above.

Unlike the AM algorithm, for MM algorithm we do not need to do alternation.

We only need to calculate the eigenvectors, which can be very fast using standard software package.

7 Simulations

The main purpose of this section is to access the adequacy of the asymptotic distributions in approximating their finite sample counterparts. To allow graphically presenting the distribution of the estimated factors and loadings, we consider the case with one factor. For $i = 1, ..., N$ and $t = 1, ..., T$, f_t and λ_i are $i.i.d.\mathcal{N}(0,1)$ and once generated, they are normalized to f_t^G and λ_i^G such that $\frac{1}{T}\sum_{t=1}^{T}(f_t^G)^2 = \frac{1}{N}\sum_{i=1}^{N}(\lambda_i^G)^2$. f_t^G and λ_i^G are fixed down for each simulation. For the given f_t^G and λ_i^G, we consider three data generating processes (DGPs) for x_{it}. Results for more DGPs, e.g. Poisson, Tobit or others, can be provided if requested.

DGP 1 (Logit): For $i = 1, ..., N$ and $t = 1, ..., T$, x_{it} is a binary random variable and $P(x_{it} = 1) = \Psi(f_t^G \lambda_i^G)$, where $\Psi(z) = 1/(1 + e^{-z})$.

DGP 2 (Probit): For $i = 1, ..., N$ and $t = 1, ..., T$, x_{it} is a binary random variable and $P(x_{it} = 1) = \Phi(f_t^G \lambda_i^G)$, where $\Phi(\cdot)$ is the cumulative distribution function of standard normal distribution.

DGP 3 (Mixed): For $i = 1, ..., 2N/5$ and $t = 1, ..., T$, x_{it} is a binary random variable and $P(x_{it} = 1) = \Psi(f_t^G \lambda_i^G)$; for $i = 2N/5 + 1, ..., 4N/5$ and $t = 1, ..., T$, x_{it} is binary random variable and $P(x_{it} = 1) = \Phi(f_t^G \lambda_i^G)$; for $i = 4N/5 + 1, ..., N$ and $t = 1, ..., T$, x_{it} is normally distributed with mean $f_t^G \lambda_i^G$ and variance 1.

Once $\{x_{it}; i = 1, ..., N, t = 1, ..., T\}$ is generated, we use the MM algorithm[8] to calculate the maximum likelihood estimators, $\{\hat{f}_t, t = 1, ..., T\}$ and $\{\hat{\lambda}_i, i = 1, ..., N\}$. For step 1, the initial values of the factors and loadings, $(\hat{f}_t^{(0)}, \hat{\lambda}_i^{(0)})$ are randomly generated from standard normal distribution for DGP1 and $Uniform(-2, 2)$ for DGP2 and DGP3[9]. For step 2, we choose $b_U = \frac{1}{4}$ for DGP1 and $b_U = 1$ for DGP2 and DGP3. This is because $-\partial_{\pi^2} l_{it}(\cdot)$ is bounded by $\frac{1}{4}$ for the Logit case, by 1 for the Probit case and equals 1 for the Gaussian case. For step 3, the maximum number

[8] We choose the MM algorithm because it is computationally simpler than the AM algorithm.

[9] We choose $U(-2, 2)$ for DGP2 and DGP3 partly because Matlab's default computational accuracy is limited.

of iteration is 20. In simulations, we find the convergence speed is very fast at the beginning. The difference between the fourth iteration and the twentieth iteration is not large. The number of simulations is 2000.

Due to limited space, we only present results for $(N, T) = (50, 50)$ and $(100, 100)$. According to Theorem 2, $N^{\frac{1}{2}}\Sigma_{t\Lambda}^{\frac{1}{2}}(\hat{f}_t - f_t^G)$ follows standard normal distribution[10] for each t and so does $T^{\frac{1}{2}}\Sigma_{iF}^{\frac{1}{2}}(\hat{\lambda}_i - \lambda_i^G)$ for each i. Figure 1 displays the histograms of $N^{\frac{1}{2}}\Sigma_{T/2,\Lambda}^{\frac{1}{2}}(\hat{f}_{T/2} - f_{T/2}^G)$ for the three DGPs. Figure 2 displays the histograms of $T^{\frac{1}{2}}\Sigma_{N/2,F}^{\frac{1}{2}}(\hat{\lambda}_{N/2} - \lambda_{N/2}^G)$ for DGP1 and DGP2. For DGP3, Figure 3 displays the histograms of $T^{\frac{1}{2}}\Sigma_{i,F}^{\frac{1}{2}}(\hat{\lambda}_i - \lambda_i^G)$ for $i = N/5$, $3N/5$ and $9N/10$. The histograms are normalized to be a density function and the standard normal density curve is overlaid on them for comparison. It is easy to see that in all subfigures, the standard normal density curve provides good approximation to the normalized histograms. Note that for different subfigures, the variance of the unnormalized estimation error, i.e., $\hat{f}_t - f_t^G$ and $\hat{\lambda}_i - \lambda_i^G$, varies with N, T and DGP of x_{it}. But once normalized, the estimation errors always approximately follow the standard normal distribution. Also, the approximation is better as N and T increases from 50 to 100. These together lend strong support to the theoretical results.

Now we consider the factor-augmented regression, $y_{t+1} = \alpha' f_t^0 + \beta' W_t + \epsilon_{t+1}$. We already have f_t^0 and $\hat{f}_t$. W_t is $i.i.d.\mathcal{N}(0, 1)$ and is fixed down once generated. $\{\epsilon_{t+1}, t = 1, ..., T\}$ is $i.i.d.\mathcal{N}(0, 1)$ and generated 2000 times. For the regression coefficients, we choose $\alpha = \beta = 1$. According to Theorem 4, $(\hat{y}_{T+1|T} - y_{T+1|T})/B_T$ should follow standard normal distribution. Figure 4 displays its histograms for the three DGPs. As Figures 1-3, the standard normal density curve is overlaid on the normalized histograms. On the whole, standard normal distribution provides reasonable approximation. The slight skewness of the histograms for the Logit case disappears if we further increase N and T. Theorem 4 also allows constructing confidence intervals for the conditional mean $y_{T+1|T}$ and the one step ahead forecast. The 95% confidence interval is $(\hat{y}_{T+1|T} - 1.96B_T, \hat{y}_{T+1|T} + 1.96B_T)$ for $y_{T+1|T}$ and

[10]Note that here $\Sigma_{iF} = \Omega_{iF}$, $\Sigma_{t\Lambda} = \Omega_{t\Lambda}$, and since f_t and λ_i are $i.i.d.\mathcal{N}(0, 1)$ and $N = T$, we have $\bar{G} = 1$.

Table 1: Coverage Rates of Confidence Intervals

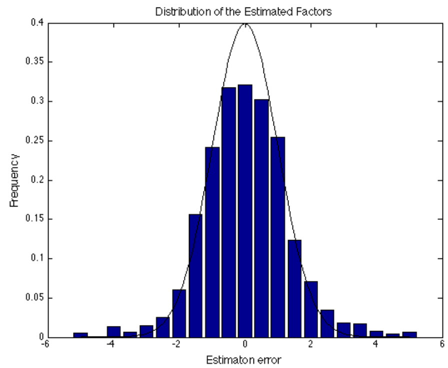
		Logit		Probit		Mixed	
N	T	$\hat{y}_{T+h\mid T}$	$\hat{y}_{T+h}$	$\hat{y}_{T+h\mid T}$	$\hat{y}_{T+h}$	$\hat{y}_{T+h\mid T}$	$\hat{y}_{T+h}$
50	50	0.954	0.947	0.946	0.948	0.959	0.950
50	100	0.955	0.951	0.961	0.950	0.943	0.952
100	50	0.931	0.943	0.961	0.951	0.954	0.952
100	100	0.962	0.944	0.941	0.950	0.948	0.951

$(\hat{y}_{T+1|T} - 1.96\sqrt{B_T^2 + \sigma_\epsilon^2}, \hat{y}_{T+1|T} + 1.96\sqrt{B_T^2 + \sigma_\epsilon^2})$ for the one step ahead forecast. Table 1 reports the coverage rates for the three DGPs. In all cases, the coverage rate is close to the nominal level 95%.

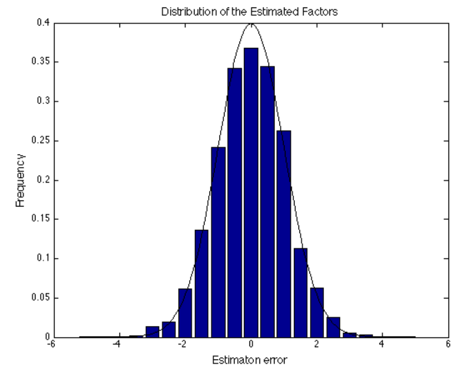
8 Conclusions

This paper studies maximum likelihood estimation of factor models with high dimensional nonlinear/mixed data. Convergence rates of the estimated factor space and loading space and asymptotic normality of the estimated factors and loadings are established under mild conditions that allows for linear models, Logit, Probit, Tobit, Poisson and some other nonlinear models. This paper also establishes the limit distributions of the parameter estimates, the conditional mean as well as the forecast when these estimated factors are used as proxies in factor-augmented regressions. These results provide a rigorous treatment of high dimensional nonlinear/mixed data in factor analysis and factor-augmented regressions. Given the prevalence of nonlinear/mixed data, empirical applications of the results developed in this paper should be fairly fruitful, especially to the topics discussed in the Introduction. For example, it would be interesting to apply this paper's method to real credit default data. We hope this paper would trigger further developments in the analysis of high dimensional nonlinear data.

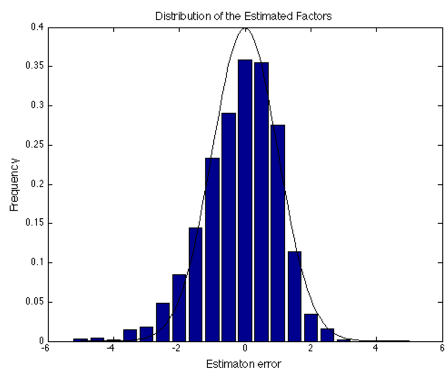
Figure 1: Distribution of the Estimated Factors



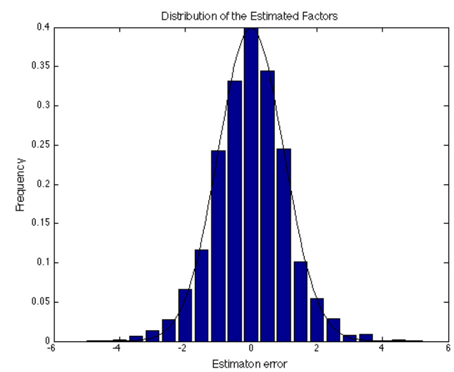
Logit, $N = 50, T = 50$.



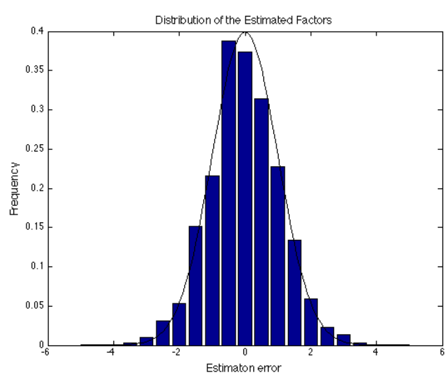
Logit, $N = 100, T = 100$.



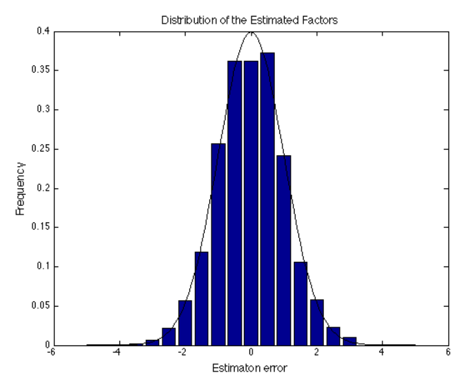
Probit, $N = 50, T = 50$.



Probit, $N = 100, T = 100$.



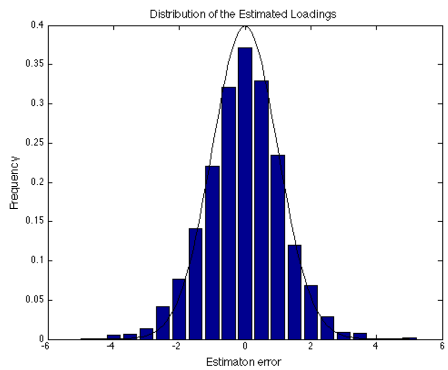
Mixed, $N = 50, T = 50$.



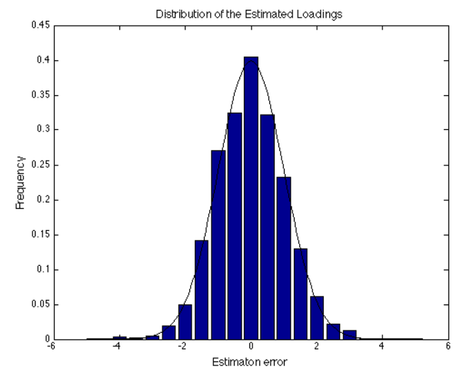
Mixed, $N = 100, T = 100$.

Notes: These histograms are for the standardized estimated factors. The curve overlaid on the histograms is the standard normal density function.

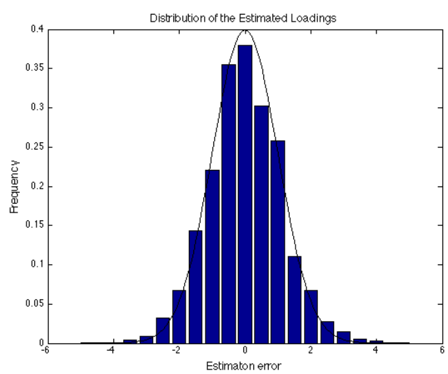
Figure 2: Distribution of the Estimated Loadings (Logit and Probit)



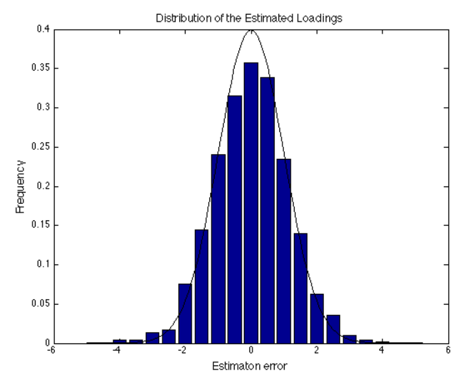
Logit, $N = 50$, $T = 50$.



Logit, $N = 100, T = 100$.



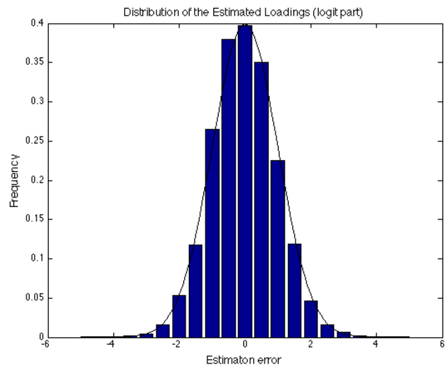
Probit, $N = 50$, $T = 50$.



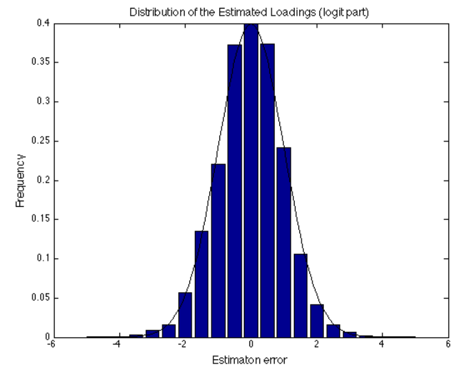
Probit, $N = 100, T = 100$.

Notes: These histograms are for the standardized estimated loadings. The curve overlaid on the histograms is the standard normal density function.

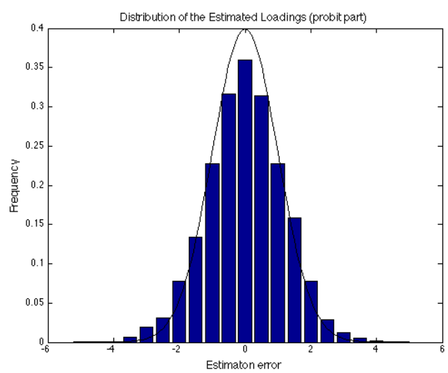
Figure 3: Distribution of the Estimated Loadings (Mixed)



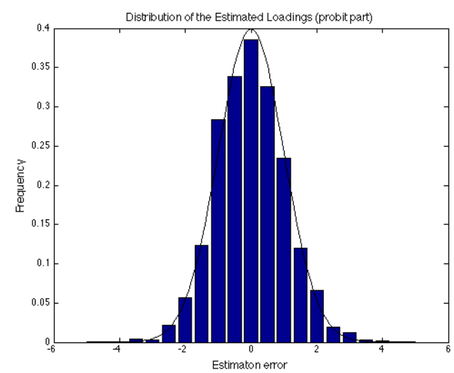
Mixed (logit part), $N = 50, T = 50$.



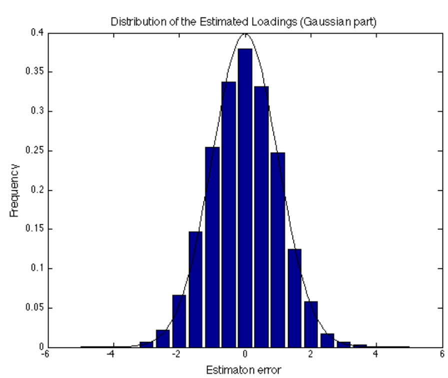
Mixed (logit part), $N = 100, T = 100$.



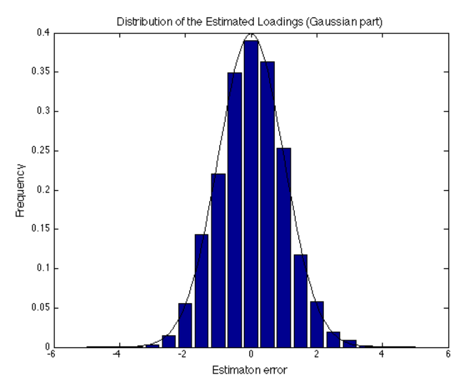
Mixed (probit part), $N = 50, T = 50$.



Mixed (probit part), $N = 100, T = 100$.



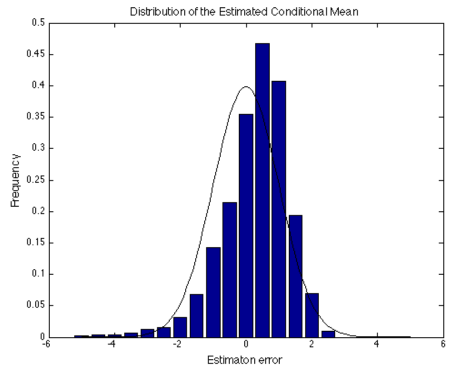
Mixed (Gaussian part), $N = 50, T = 50$.



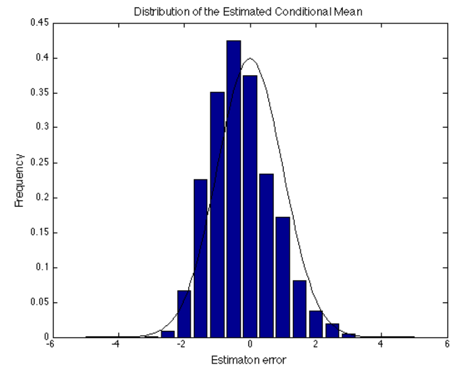
Mixed (Gaussian part), $N = 100, T = 100$.

Notes: These histograms are for the standardized estimated loadings. The curve overlaid on the histograms is the standard normal density function.

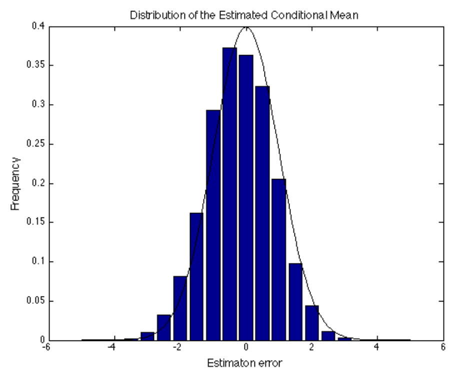
Figure 4: Distribution of the Estimated Conditional Mean



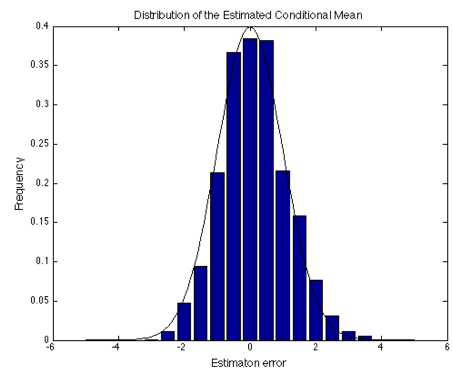
Logit, $N = 50$, $T = 50$.



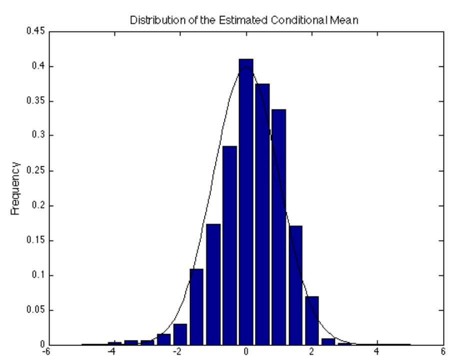
Logit, $N = 100, T = 100$.



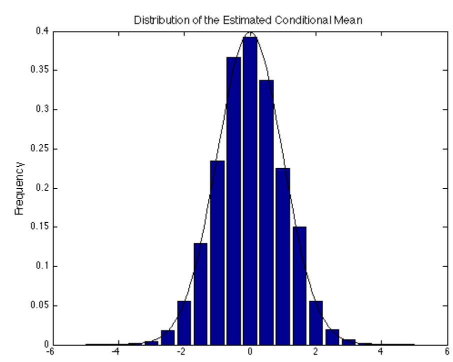
Probit, $N = 50, T = 50$.



Probit, $N = 100, T = 100$.



Mixed, $N = 50, T = 50$.



Mixed, $N = 100, T = 100$.

Notes: These histograms are for the standardized estimated conditional mean. The curve overlaid on the histograms is the standard normal density function.

References

[1] Bai (2003): "Inferential Theory for Factor Models of Large Dimensions," *Econometrica*, 71, 135-171.

[2] Bai (2009): "Panel Data Models with Interactive Fixed Effects," *Econometrica*, 77, 1229-1279.

[3] Bai and Li (2012): "Statistical Analysis of Factor Models of High Dimension," *Annals of Statistics,* 436-465.

[4] Bai and Li (2016): "Maximum Likelihood Estimation and Inference for Approximate Factor Models of High Dimension," *Review of Economics and Statistics,* 98, 298-309.

[5] Bai and Ng (2002): "Determining the Number of Factors in Approximate Factor Models," *Econometrica*, 70, 191–221.

[6] Bai and Ng (2006): "Confidence Intervals for Diffusion Index Forecasts and Inference for Factor-augmented Regressions," *Econometrica*, 74, 1133-1150.

[7] Bartholomew (1980): "Factor Analysis for Categorical Data," *Journal of the Royal Statistical Society. Series B*, 293-321.

[8] Bartholomew and Knott (1999): Latent Variable Models and Factor Analysis. Edward Arnold.

[9] Bernanke, Boivin and Eliasz (2005): "Measuring the Effects of Monetary Policy: A Factor-augmented Vector Autoregressive (FAVAR) approach," *Quarterly Journal of Economics*, 120, 387–422.

[10] Bohning and Lindsay (1988): "Monotonicity of Quadratic-approximation Algorithms," *Annals of the Institute of Statistical Mathematics*, 40, 641–663.

[11] Campbell, Lo and Mackinlay (1997): The Econometrics of Financial Markets. New Jersey: Princeton University Press.

[12] Chen (2016): "Estimation of Nonlinear Panel Models with Multiple Unobserved Effects," Warwick Economics Research Paper Series.

[13] Chen, Fernandez-Val and Weidner (2014): "Nonlinear Panel Models with Interactive Effects," arXiv preprint arXiv:1412.5647.

[14] Chen, Fernandez-Val and Weidner (2018): "Nonlinear Factor Models for Network and Panel Data," Cemmap Working Paper CWP38/18

[15] Collins, Dasgupta and Schapire (2001): "A Generalization of Principal Component Analysis to the Exponential Family," *Advances in Neural Information Processing System,* Vol. 13.

[16] Cox and Reid (1987): "Parameter Orthogonality and Approximate Conditional Inference," *Journal of the Royal Statistical Society. Series B*, 1-39.

[17] Creal, Schwaab, Koopman and Lucas (2014): "Observation-driven Mixed-measurement Dynamic Factor Models with an Application to Credit Risk," *Review of Economics and Statistics*, 96, 898-915.

[18] de Leeuw (2006): "Principal Component Analysis of Binary Data by Iterated Singular Value Decomposition," *Computational Statistics & Data Analysis,* 50, 21-39.

[19] Feng, Wang, Han, Xia and Tu (2013): "The Mean Value Theorem and Taylor's Expansion in Statistics," *American Statistician*, 67, 245-248.

[20] Fernandez-Val and Weidner (2016): "Individual and Time Effects in Nonlinear Panel Models with Large N, T," *Journal of Econometrics,* 192, 291-312.

[21] Filmer and Pritchett (2001): "Estimating Wealth Effects without Expenditure Data—or Tears: An Application to Educational Enrollments in States of India," *Demography*, 38, 115-132.

[22] Hahn and Newey (2004): "Jackknife and Analytical Bias Reduction for Nonlinear Panel Models," *Econometrica*, 72, 1295–1319.

[23] Hahn and Kuersteiner (2011): "Bias Reduction for Dynamic Nonlinear Panel Models with Fixed Effects," *Econometric Theory*, 27, 1152-1191.

[24] Hunter and Lange (2004): "A Tutorial on MM Algorithms," *American Statistician*, 58, 30-37.

[25] Koopman and Lucas (2008): "A Non-Gaussian Panel Time Series Model for Estimating and Decomposing Default Risk," *Journal of Business & Economic Statistics,* 26, 510-525.

[26] Koopman, Lucas and Monteiro (2008): "The Multi-state Latent Factor Intensity Model for Credit Rating Transitions," *Journal of Econometrics,* 142, 399-424.

[27] Koopman, Lucas and Schwaab (2011): "Modeling Frailty-correlated Defaults Using Many Macroeconomic Covariates," *Journal of Econometrics,* 162, 312-325.

[28] Jennrich (1969): "Asymptotic Properties of Non-linear Least Squares Estimators," *The Annals of Mathematical Statistics*, 40, 633-643.

[29] Joreskog and Moustaki (2001): "Factor Analysis of Ordinal Variables: A Comparison of Three Approaches," *Multivariate Behavioral Research*, 36, 347-387.

[30] Lancaster (2000): "The Incidental Parameter Problem since 1948," *Journal of Econometrics*, 95, 391-413.

[31] Lancaster (2002): "Orthogonal Parameters and Panel Data," *Review of Economic Studies*, 69, 647-666.

[32] Lange, Hunter and Young (2000): "Optimization Transfer Using Surrogate Objective Functions," *Journal of Computational and Graphical Statistics*, 9, 1-20.

[33] McNeil and Wendin (2007): "Bayesian Inference for Generalized Linear Mixed Models of Portfolio Credit Risk," *Journal of Empirical Finance,* 14, 131-149.

[34] Moustaki (1996): "A Latent Trait and a Latent Class Model for Mixed Observed Variables," *British Journal of Mathematical and Statistical Psychology*, 49, 313-334.

[35] Moustaki (2000): "A Latent Variable Model for Ordinal Variables," *Applied Psychological Measurement*, 24, 211-223.

[36] Moustaki and Knott (2000): "Generalized Latent Trait Models," *Psychometrika*, 65, 391-411.

[37] Newey and McFadden (1994): "Large sample Estimation and Hypothesis Testing," *Handbook of Econometrics*, Vol. IV, 2111-2245.

[38] Ng (2015): "Constructing Common Factors from Continuous and Categorical Data," *Econometric Reviews*, 34, 1141-1171.

[39] Ross (1976): "The Arbitrage Theory of Capital Asset Pricing," *Journal of Finance*, 13, 341–360.

[40] Schein, Saul and Ungar (2003): "A Generalized Linear Model for Principal Component Analysis of Binary Data," *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics.*

[41] Schonbucher (2000): "Factor Models for Portfolio Credit Risk," Bonn Econ Discussion Papers 16.

[42] Stock and Watson (2002): "Forecasting Using Principal Components from a Large Number of Predictors," *Journal of American Statistical Association*, 97, 1167–1179.

[43] Stock and Watson (2016): "Factor Models and Structural Vector Autoregressions in Macroeconomics," *Handbook of Macroeconomics*, forthcoming.

[44] Wu (1981): "Asymptotic Theory of Nonlinear Least Squares Estimation," *The Annals of Statistics,* 501–513.

Maximum Likelihood Estimation and Inference for High Dimensional Nonlinear Factor Models with Application to Factor-augmented Regressions

Appendix

A Structure of the Hessian

Since $\partial_{\phi}P(\lambda^G, f^G) = 0$, the score is

$$S = (\sum_{t=1}^{T} \partial_{\pi} l_{1t} f_t^{G\prime}, ..., \sum_{t=1}^{T} \partial_{\pi} l_{Nt} f_t^{G\prime}, \sum_{i=1}^{N} \partial_{\pi} l_{i1} \lambda_i^{G\prime}, ..., \sum_{i=1}^{N} \partial_{\pi} l_{iT} \lambda_i^{G\prime})'. \quad (15)$$

For the Hessian, define

$$H_L(\phi) = \begin{bmatrix} H_{L\lambda\lambda'}(\phi) & H_{L\lambda f'}(\phi) \\ H_{Lf\lambda'}(\phi) & H_{Lff'}(\phi) \end{bmatrix}. \quad (16)$$

and

$$J_L(\phi) = \begin{bmatrix} 0 & J_{L\lambda f'}(\phi) \\ J_{Lf\lambda'}(\phi) & 0 \end{bmatrix}. \quad (17)$$

$H_{L\lambda\lambda'}(\phi)$ is of dimension $Nr \times Nr$ and block-diagonal. Each block is $r \times r$ and the i-th diagonal block is $\sum_{t=1}^{T} \partial_{\pi^2} l_{it}(\pi_{it}) f_t f_t'$. $H_{Lff'}(\phi)$ is of dimension $Tr \times Tr$ and block-diagonal. Each block is $r \times r$ and the t-th diagonal block is $\sum_{k=1}^{N} \partial_{\pi^2} l_{kt}(\pi_{kt}) \lambda_k \lambda_k'$. $H_{L\lambda f'}(\phi)$ is of dimension $Nr \times Tr$. Each block is $r \times r$ and the (i,t) block is $\partial_{\pi^2} l_{it}(\pi_{it}) f_t \lambda_i'$. $H_{Lf\lambda'}(\phi)$ is the transpose of $H_{L\lambda f'}(\phi)$. $J_{L\lambda f'}(\phi)$ is of dimension $Nr \times Tr$. Each block is $r \times r$ and the (i,t) block is $\partial_{\pi} l_{it}(\pi_{it}) I_r$. $J_{Lf\lambda'}(\phi)$ is the transpose of $J_{L\lambda f'}(\phi)$. It follows that

$$\partial_{\phi\phi'} L(\phi) = H_L(\phi) + J_L(\phi). \quad (18)$$

Let $H_P(\phi) = \partial_{\phi\phi'} P(\phi)$, then

$$H(\phi) = H_L(\phi) + J_L(\phi) + H_P(\phi). \quad (19)$$

Next, for $p = 1, ..., r$ and $q = p + 1, ..., r$, define v_p, u_{pq} and u_{qp} as follows. Let v_p be a $Nr + Tr$ dimensional vector. For the first Nr elements, in the i-th block, the

p-th element is λ_{ip} and all the other elements are zeros. For the last Tr elements, in the t-th block, the p-th element is $-f_{tp}$ and all the other elements are zeros. Let u_{pq} be a $Nr+Tr$ dimensional vector. The last Tr elements are all zeros. For the first Nr elements, in the i-th block, the p-th element is λ_{iq}, the q-th element is λ_{ip} and all the other elements are zeros. Let u_{qp} be a $Nr+Tr$ dimensional vector. The first Nr elements are all zeros. For the last Tr elements, in the t-th block, the p-th element is f_{tq}, the q-th element is f_{tp} and all the other elements are zeros. Also, when $\lambda=\lambda^G$ and $f=f^G$, v_p, u_{pq} and u_{qp} are denoted as v_p^G, u_{pq}^G and u_{qp}^G respectively. It can be verified that

$$\begin{aligned}
\partial_{\phi}[(\frac{\sum_{i=1}^N\lambda_{iq}^2}{N}-\frac{\sum_{t=1}^T f_{tq}^2}{T})^2] &= 4(\frac{\sum_{i=1}^N\lambda_{iq}^2}{N}-\frac{\sum_{t=1}^T f_{tq}^2}{T})D_{NT}^{-1}v_q, \\
\partial_{\phi\phi'}[(\frac{\sum_{i=1}^N\lambda_{iq}^2}{N}-\frac{\sum_{t=1}^T f_{tq}^2}{T})^2] &= 8D_{NT}^{-1}v_q v_q' D_{NT}^{-1} \\
&\quad +4(\frac{\sum_{i=1}^N\lambda_{iq}^2}{N}-\frac{\sum_{t=1}^T f_{tq}^2}{T})D_{NT}^{-1}(I_{N+T}\otimes\iota_q), \quad (20)
\end{aligned}$$

where ι_q is an $r\times r$ matrix with the q-th diagonal element being one and all the other elements being zero. Also,

$$\partial_{\phi\phi'}[(\sum_{i=1}^N\lambda_{ip}\lambda_{iq})^2] = 2[u_{pq}u_{pq}'+(\sum_{i=1}^N\lambda_{ip}\lambda_{iq})D_1], \quad (21)$$

$$\partial_{\phi\phi'}[(\sum_{t=1}^T f_{tp}f_{tq})^2] = 2[u_{qp}u_{qp}'+(\sum_{t=1}^T f_{tp}f_{tq})D_2], \quad (22)$$

where $D_1=\begin{bmatrix} I_N\otimes\iota_{pq} & 0 \\ 0 & 0\end{bmatrix}$, $D_2=\begin{bmatrix} 0 & 0 \\ 0 & I_T\otimes\iota_{pq}\end{bmatrix}$ and ι_{pq} is an $r\times r$ matrix with the (p,q) element and the (q,p) element being one and all the other elements being zero. Since $\frac{1}{T}F^{G\prime}F^G=\frac{1}{N}\Lambda^{G\prime}\Lambda^G$ and both are diagonal, the second term on the right hand side of (20), (21) and (22) are all zero when f^G and λ^G are plugged in. Thus

$$\begin{aligned}
H_P &= -c(\sum_{p=1}^r NTD_{NT}^{-1}v_p^G v_p^{G\prime}D_{NT}^{-1} \\
&\quad +\frac{T}{N}\sum_{p=1}^r\sum_{q=p+1}^r u_{pq}^G u_{pq}^{G\prime}+\frac{N}{T}\sum_{p=1}^r\sum_{q=p+1}^r u_{qp}^G u_{qp}^{G\prime}). \quad (23)
\end{aligned}$$

B Lemmas and Their Proofs

Lemma 1 *Under Assumptions 2 and 3,* $\|\partial_\pi l\| = O_p(N^{\frac{1}{4}}T^{\frac{1}{2}} + N^{\frac{1}{2}}T^{\frac{1}{4}})$.

Proof. Let $\partial_\pi l$ denote the $N \times T$ matrix with $\partial_\pi l_{it}$ in the i-th row and t-th column. We shall show $\mathbb{E}\,\|\partial_\pi l\|^4 = O(NT^2 + N^2T)$. First note that

$$\|\partial_\pi l\|^4 = \|(\partial_\pi l)'\partial_\pi l\|^2 \leq \|(\partial_\pi l)'\partial_\pi l\|_F^2 = \sum\nolimits_{s=1}^{T}\sum\nolimits_{t=1}^{T}(\sum\nolimits_{i=1}^{N}\partial_\pi l_{is}\partial_\pi l_{it})^2.$$

It is easy to see that $\mathbb{E}(\sum_{i=1}^N \partial_\pi l_{is}\partial_\pi l_{it})^2$ is not larger than the sum of $2\mathbb{E}(\sum_{i=1}^N[\partial_\pi l_{is}\partial_\pi l_{it} - \mathbb{E}(\partial_\pi l_{is}\partial_\pi l_{it})])^2$ and $2(\sum_{i=1}^N \mathbb{E}(\partial_\pi l_{is}\partial_\pi l_{it}))^2$. Under Assumption 3, the former is not larger than NT^2M while the latter is not larger than N^2TM.

Note that the order $O_p(N^{\frac{1}{4}}T^{\frac{1}{2}} + N^{\frac{1}{2}}T^{\frac{1}{4}})$ is not sharp. Results in random matrix theory show that if $\partial_\pi l_{it}$ is independent over i and t and its fourth moment is uniformly bounded over i and t, then $\|\partial_\pi l\| = O_p(\max\{N^{\frac{1}{2}}, T^{\frac{1}{2}}\})$. But random matrix theory has not established this result under weak dependence over i and t. Lemma 1 allows for serial and cross-sectional dependence. Although not sharp, its order is enough for proving Proposition 1. ■

Lemma 2 *Under Assumptions 1, 2 and 6,* $\left\|(D_{TN}^{-\frac{1}{2}}HD_{TN}^{-\frac{1}{2}})^{-1}\right\| = O_p(1)$ *as* $(N,T) \to \infty$.

Proof. Step (1): We first introduce w_{pq}^G and w_{qp}^G. For $p = 1, ..., r$ and $q = p+1, ..., r$, w_{pq}^G is a $Nr + Tr$ dimensional vector. For the first Nr elements, in the i-th block, the p-th element is λ_{iq}^G and all the other elements are zero. For the last Tr elements, in the t-th block, the q-th element is $-f_{tp}^G$ and all the other elements are zero. w_{qp}^G is a $Nr + Tr$ dimensional vector. For the first Nr elements, in the i-th block, the q-th element is λ_{ip}^G and all the other elements are zero. For the last Tr elements, in the t-th block, the p-th element is $-f_{tq}^G$ and all the other elements are zero. It can be verified that under condition (4),

1. for $p = 1, ..., r$ and $q = p+1, ..., r$, $D_{NT}^{-\frac{1}{2}}v_p^G$, $D_{NT}^{-\frac{1}{2}}w_{pq}^G$ and $D_{NT}^{-\frac{1}{2}}w_{qp}^G$ are all orthogonal to the space spanned by eigenvectors of $D_{TN}^{-\frac{1}{2}}H_L D_{TN}^{-\frac{1}{2}}$,

2. for any $p_1 = 1, ..., r$, $p_2 = 1, ..., r$, $p_3 = 1, ..., r$, $q_2 = p_2 + 1, ..., r$ and $q_3 = p_3 + 1, ..., r$, $D_{NT}^{-\frac{1}{2}} v_{p_1}^G$, $D_{NT}^{-\frac{1}{2}} w_{p_2 q_2}^G$ and $D_{NT}^{-\frac{1}{2}} w_{q_3 p_3}^G$ are orthogonal to each other.

Let $V = (v_1^G, ..., v_r^G) = (V_\lambda', V_f')'$ and

$$\begin{aligned} W &= (w_{12}^G, ..., w_{1r}^G, w_{23}^G, ..., w_{2r}^G, ..., w_{(r-1)r}^G; w_{21}^G, ..., w_{r1}^G, w_{32}^G, ..., w_{r2}^G, ..., w_{r(r-1)}^G) \\ &= (W_\lambda', W_f')'. \end{aligned}$$

Note that V, V_λ, V_f, W, W_λ and W_f are of dimension $(Nr + Tr) \times r$, $Nr \times r$, $Tr \times r$,$(Nr + Tr) \times r(r-1)$, $Nr \times r(r-1)$, $Tr \times r(r-1)$ respectively. Next, define $\breve{H}$ such that

$$\begin{aligned} D_{TN}^{-\frac{1}{2}} \breve{H} D_{TN}^{-\frac{1}{2}} &= D_{TN}^{-\frac{1}{2}} H_L D_{TN}^{-\frac{1}{2}} - c D_{NT}^{-\frac{1}{2}} (\sum\nolimits_{p=1}^r v_p^G v_p^{G\prime} + \\ &\quad \sum\nolimits_{p=1}^r \sum\nolimits_{q=p+1}^r w_{pq}^G w_{pq}^{G\prime} + \sum\nolimits_{p=1}^r \sum\nolimits_{q=p+1}^r w_{qp}^G w_{qp}^{G\prime}) D_{NT}^{-\frac{1}{2}}, \end{aligned}$$

and let $\breve{H}_{\lambda\lambda'}$, $\breve{H}_{\lambda f'}$, $\breve{H}_{f\lambda'}$ and $\breve{H}_{ff'}$ be the upper-left, upper-right, lower-left and lower-right block of $\breve{H}$. It can be verified that $\breve{H}_{\lambda f'}$ is of dimension $Nr \times Tr$ and the (i, t) block is $(\partial_{\pi^2} l_{it} + c) \times f_t^G \lambda_i^{G\prime}$, and

$$\breve{H}_{\lambda\lambda'} = H_{L\lambda\lambda'} - c\frac{T}{N}(V_\lambda, W_\lambda)(V_\lambda, W_\lambda)', \tag{24}$$

$$\breve{H}_{ff'} = H_{Lff'} - c\frac{N}{T}(V_f, W_f)(V_f, W_f)'. \tag{25}$$

$-D_{TN}^{-\frac{1}{2}} \breve{H} D_{TN}^{-\frac{1}{2}}$ can be written as

$$D_{TN}^{-\frac{1}{2}} [\sum\nolimits_{i=1}^N \sum\nolimits_{t=1}^T (-\partial_{\pi^2} l_{it} - c) \begin{bmatrix} 1_i^{(N)} \otimes f_t^G \\ 1_t^{(T)} \otimes \lambda_i^G \end{bmatrix} \begin{bmatrix} 1_i^{(N)} \otimes f_t^G \\ 1_t^{(T)} \otimes \lambda_i^G \end{bmatrix}' D_{TN}^{-\frac{1}{2}} \tag{26}$$

$$+ \begin{bmatrix} I_N \otimes c\frac{1}{T} \sum_{t=1}^T f_t^G f_t^{G\prime} & 0 \\ 0 & I_T \otimes c\frac{1}{N} \sum_{i=1}^N \lambda_i^G \lambda_i^{G\prime} \end{bmatrix} \tag{27}$$

$$+ \begin{bmatrix} c\frac{1}{N}(V_\lambda, W_\lambda)(V_\lambda, W_\lambda)' & 0 \\ 0 & c\frac{1}{T}(V_f, W_f)(V_f, W_f)' \end{bmatrix}, \tag{28}$$

where $1_x^{(y)}$ is an y dimensional vector with the x-th element being one and all the other elements being zero. Expressions (26) and (28) are positive semi-definite.

Since $\frac{1}{T}\sum_{t=1}^{T} f_t^G f_t^{G\prime} = \frac{1}{N}\sum_{i=1}^{N}\lambda_i^G\lambda_i^{G\prime} = \mathcal{V}^{\frac{1}{2}}$ and $\mathcal{V}$ converges in probability to the diagonal matrix of eigenvalues of $\Sigma_\Lambda^{\frac{1}{2}}\Sigma_F\Sigma_\Lambda^{\frac{1}{2}}$, the minimum eigenvalue of (27) is positive and bounded away from zero w.p.a.1. Thus there exists some $b > 0$ such that $\rho_{\min}(-D_{TN}^{-\frac{1}{2}}\breve{H}D_{TN}^{-\frac{1}{2}}) \geq b$ w.p.a.1.

Step (2): The positive definiteness of $-D_{TN}^{-\frac{1}{2}}\breve{H}D_{TN}^{-\frac{1}{2}}$ implies that eigenvectors of $D_{TN}^{-\frac{1}{2}}H_L D_{TN}^{-\frac{1}{2}}$ together with $\{\frac{D_{NT}^{-\frac{1}{2}}v_p^G}{\left\|D_{NT}^{-\frac{1}{2}}v_p^G\right\|}, p = 1, ..., r\}$, $\{\frac{D_{NT}^{-\frac{1}{2}}w_{pq}^G}{\left\|D_{NT}^{-\frac{1}{2}}w_{pq}^G\right\|}, p = 1, ..., r, q = p + 1, ..., r\}$ and $\{\frac{D_{NT}^{-\frac{1}{2}}w_{qp}^G}{\left\|D_{NT}^{-\frac{1}{2}}w_{qp}^G\right\|}, p = 1, ..., r, q = p+1, ..., r\}$ constitutes an orthonormal basis. Under this basis, for $j = 1, ..., r$ and $k = j+1, ..., r$, let $(u_{jk,1}^G, ..., u_{jk,(N+T)r-r(r-1)}^G)$ be the coordinates of u_{jk}^G corresponding to eigenvectors of $D_{TN}^{-\frac{1}{2}}H_L D_{TN}^{-\frac{1}{2}}$ and $\frac{D_{NT}^{-\frac{1}{2}}v_p^G}{\left\|D_{NT}^{-\frac{1}{2}}v_p^G\right\|}$, and let $u_{jk,pq}^G$ and $u_{jk,qp}^G$ be the coordinate of u_{jk}^G corresponding to $\frac{D_{NT}^{-\frac{1}{2}}w_{pq}^G}{\left\|D_{NT}^{-\frac{1}{2}}w_{pq}^G\right\|}$ and $\frac{D_{NT}^{-\frac{1}{2}}w_{qp}^G}{\left\|D_{NT}^{-\frac{1}{2}}w_{qp}^G\right\|}$ respectively. Coordinates of u_{kj}^G are defined in the same way.

To prove the Lemma, it suffices to show that there exists $C > 0$ such that for any vector a with $\|a\| = 1$, $a'(-D_{TN}^{-\frac{1}{2}}HD_{TN}^{-\frac{1}{2}})a \geq C > 0$ w.p.a.1 as $(N, T) \to \infty$. Let

$$(a_1, ..., a_{(N+T)r-r(r-1)}; a_{12}, ..., a_{1r}, a_{23}, ..., a_{2r}, ..., a_{(r-1)r}; a_{21}, ..., a_{r1}, a_{32}, ..., a_{r2}, ..., a_{r(r-1)})$$

be the coordinates of a. Plug in equations (19) and (23), we have

$$\begin{aligned} a'(-D_{TN}^{-\frac{1}{2}}HD_{TN}^{-\frac{1}{2}})a &= a'(-D_{TN}^{-\frac{1}{2}}H_L D_{TN}^{-\frac{1}{2}} + c\sum_{p=1}^{r} D_{NT}^{-\frac{1}{2}}v_p^G v_p^{G\prime} D_{NT}^{-\frac{1}{2}})a \\ &\quad + ca'[\sum_{j=1}^{r}\sum_{k=j+1}^{r}(\frac{u_{jk}^G u_{jk}^{G\prime}}{N} + \frac{u_{kj}^G u_{kj}^{G\prime}}{T})]a \\ &\quad - N^{-\frac{1}{2}}T^{-\frac{1}{2}}a'J_L a. \end{aligned} \tag{29}$$

The first term on the right hand side of (29) is not smaller than $b\sum_{l=1}^{(N+T)r-r(r-1)} a_l^2$ w.p.a.1 because the smallest nonzero eigenvalue of $-D_{TN}^{-\frac{1}{2}}H_L D_{TN}^{-\frac{1}{2}} + c\sum_{p=1}^{r} D_{NT}^{-\frac{1}{2}}v_p^G v_p^{G\prime} D_{NT}^{-\frac{1}{2}}$ is not smaller than $\rho_{\min}(-D_{TN}^{-\frac{1}{2}}\breve{H}D_{TN}^{-\frac{1}{2}})$. The second term on the right hand side of (29) is not smaller than $c_1\sum_{j=1}^{r}\sum_{k=j+1}^{r}[\frac{(a'u_{jk}^G)^2}{N} + \frac{(a'u_{kj}^G)^2}{T}]$ for some $0 < c_1 < c$. How to choose c_1 will be discussed later. For $a'u_{jk}^G$, we have

$$\begin{aligned}
(a'u_{jk}^G)^2 &= [\sum_{l=1}^{(N+T)r-r(r-1)} a_l u_{jk,l}^G + \sum_{p=1}^{r}\sum_{q=p+1}^{r}(a_{pq}u_{jk,pq}^G + a_{qp}u_{jk,qp}^G)]^2 \\
&= (\sum_{l=1}^{(N+T)r-r(r-1)} a_l u_{jk,l}^G)^2 + [\sum_{p=1}^{r}\sum_{q=p+1}^{r}(a_{pq}u_{jk,pq}^G + a_{qp}u_{jk,qp}^G)]^2 \\
&\quad +2(\sum_{l=1}^{(N+T)r-r(r-1)} a_l u_{jk,l}^G)[\sum_{p=1}^{r}\sum_{q=p+1}^{r}(a_{pq}u_{jk,pq}^G + a_{qp}u_{jk,qp}^G)] \\
&\geq [\sum_{p=1}^{r}\sum_{q=p+1}^{r}(a_{pq}u_{jk,pq}^G + a_{qp}u_{jk,qp}^G)]^2 - 2(\sum_{l=1}^{(N+T)r-r(r-1)} a_l^2)^{\frac{1}{2}} \left\|u_{jk}^G\right\|^2,
\end{aligned}$$

where the last inequality follows from $\left|\sum_{p=1}^{r}\sum_{q=p+1}^{r}(a_{pq}u_{jk,pq}^G + a_{qp}u_{jk,qp}^G)\right| \leq \|a\| \left\|u_{jk}^G\right\|$. Similarly, for $a'u_{kj}^G$, we have

$$(a'u_{kj}^G)^2 \geq [\sum_{p=1}^{r}\sum_{q=p+1}^{r}(a_{pq}u_{kj,pq}^G + a_{qp}u_{kj,qp}^G)]^2 - 2(\sum_{l=1}^{(N+T)r-r(r-1)} a_l^2)^{\frac{1}{2}} \left\|u_{kj}^G\right\|^2.$$

Thus the second term on the right hand side of (29) is not smaller than

$$\begin{aligned}
& c_1 \sum_{j=1}^{r}\sum_{k=j+1}^{r}\{\frac{1}{N}[\sum_{p=1}^{r}\sum_{q=p+1}^{r}(a_{pq}u_{jk,pq}^G + a_{qp}u_{jk,qp}^G)]^2 \\
& +\frac{1}{T}[\sum_{p=1}^{r}\sum_{q=p+1}^{r}(a_{pq}u_{kj,pq}^G + a_{qp}u_{kj,qp}^G)]^2\} \qquad (30) \\
& -2c_1(\sum_{l=1}^{(N+T)r-r(r-1)} a_l^2)^{\frac{1}{2}} \sum_{j=1}^{r}\sum_{k=j+1}^{r}(\frac{1}{N}\left\|u_{jk}^G\right\|^2 + \frac{1}{T}\left\|u_{kj}^G\right\|^2). \qquad (31)
\end{aligned}$$

By Assumption 1, expression (31) is not smaller than $-2c_1 r(r-1)M(\sum_{l=1}^{(N+T)r-r(r-1)} a_l^2)^{\frac{1}{2}}$ for some $0 < M < \infty$ w.p.a.1. To evaluate expression (30), let

$$u_{jk}^* = (u_{jk,12}^G, ..., u_{jk,1r}^G; u_{jk,23}^G, ..., u_{jk,2r}^G; ...; u_{jk,(r-2)(r-1)}^G, u_{jk,(r-2)r}^G; u_{jk,(r-1)r}^G)'$$

and define u_{kj}^* similarly. Let

$$\begin{aligned}
U^* &= [N^{-\frac{1}{2}}(u_{12}^*, ..., u_{1r}^*; u_{23}^*, ..., u_{2r}^*; ...; u_{(r-2)(r-1)}^*, u_{(r-2)r}^*; u_{(r-1)r}^*); \\
&\quad T^{-\frac{1}{2}}(u_{21}^*, ..., u_{r1}^*; u_{32}^*, ..., u_{r2}^*; ...; u_{(r-1)(r-2)}^*, u_{r(r-2)}^*; u_{r(r-1)}^*)],
\end{aligned}$$

then expression (30) is not smaller than $c_1\rho_{\min}(U^*U^{*\prime})\sum_{p=1}^{r}\sum_{q=p+1}^{r}(a_{pq}^2 + a_{qp}^2)$. Under Assumptions 1 and 6, $plimU^*$ is full rank (We shall prove this later). Thus $plimU^*U^{*\prime}$ is positive definite. This implies that there exists $d > 0$ such that

$\rho_{\min}(U^*U^{*\prime}) \geq d$ w.p.a.1 as $(N,T) \to \infty$. It follows that expression (30) is not smaller than $c_1 d \sum_{p=1}^{r} \sum_{q=p+1}^{r} (a_{pq}^2 + a_{qp}^2)$ w.p.a.1.

The third term on the right hand side of (29) is $O_p(N^{-\frac{1}{4}} + T^{-\frac{1}{4}})$. This follows from Lemma 1 and $N^{-\frac{1}{2}}T^{-\frac{1}{2}}a'J_L a \leq 2N^{-\frac{1}{2}}T^{-\frac{1}{2}} \|J_{L\lambda f'}\| \leq 2N^{-\frac{1}{2}}T^{-\frac{1}{2}} \|\partial_\pi l\|$. Thus $N^{-\frac{1}{2}}T^{-\frac{1}{2}}a'J_L a \leq \frac{c_1 d}{3}$ w.p.a.1. The above analysis together shows that w.p.a.1,

$$\begin{aligned}
& a'(-D_{TN}^{-\frac{1}{2}} H D_{TN}^{-\frac{1}{2}})a \\
\geq\ & b \sum_{l=1}^{(N+T)r-r(r-1)} a_l^2 + c_1 d \sum_{p=1}^{r} \sum_{q=p+1}^{r} (a_{pq}^2 + a_{qp}^2) \\
& -2c_1 r(r-1) M (\sum_{l=1}^{(N+T)r-r(r-1)} a_l^2)^{\frac{1}{2}} - \frac{c_1 d}{3} \\
=\ & (b - c_1 d) \sum_{l=1}^{(N+T)r-r(r-1)} a_l^2 + c_1 d - 2c_1 r(r-1) M (\sum_{l=1}^{(N+T)r-r(r-1)} a_l^2)^{\frac{1}{2}} - \frac{c_1 d}{3} \\
\geq\ & c_1 d - \frac{c_1^2 r^2 (r-1)^2 M^2}{b - c_1 d} - \frac{c_1 d}{3} = \frac{c_1(bd - c_1 d^2 - c_1 r^2 (r-1)^2 M^2)}{b - c_1 d} - \frac{c_1 d}{3}. \qquad (32)
\end{aligned}$$

When c_1 is small enough, $c_1 d^2 - c_1 r^2 (r-1)^2 M^2$ is smaller than $\frac{bd}{2}$. Thus when c_1 is small enough, the last term of expression (32) is not smaller than $\frac{c_1 d}{6}$. Take $C = \frac{c_1 d}{6}$, we have proved that $a'(-D_{TN}^{-\frac{1}{2}} H D_{TN}^{-\frac{1}{2}})a \geq C$ w.p.a.1.

Now we prove the full rankness of $plimU^*$. We shall prove for the case $r = 3$, other cases can be shown similarly. When $r = 3$, after some calculation, U^* equals

$$\begin{matrix}
\frac{\frac{1}{N}\sum_{i=1}^{N}(\lambda_{i2}^G)^2}{\left\|D_{NT}^{-\frac{1}{2}}w_{12}^G\right\|} & 0 & 0 & \frac{-\frac{1}{T}\sum_{t=1}^{T}(f_{t1}^G)^2}{\left\|D_{NT}^{-\frac{1}{2}}w_{12}^G\right\|} & 0 & 0 \\
0 & \frac{\frac{1}{N}\sum_{i=1}^{N}(\lambda_{i3}^G)^2}{\left\|D_{NT}^{-\frac{1}{2}}w_{13}^G\right\|} & 0 & 0 & \frac{-\frac{1}{T}\sum_{t=1}^{T}(f_{t1}^G)^2}{\left\|D_{NT}^{-\frac{1}{2}}w_{13}^G\right\|} & 0 \\
0 & 0 & \frac{\frac{1}{N}\sum_{i=1}^{N}(\lambda_{i3}^G)^2}{\left\|D_{NT}^{-\frac{1}{2}}w_{23}^G\right\|} & 0 & 0 & \frac{-\frac{1}{T}\sum_{t=1}^{T}(f_{t2}^G)^2}{\left\|D_{NT}^{-\frac{1}{2}}w_{23}^G\right\|} \\
\frac{\frac{1}{N}\sum_{i=1}^{N}(\lambda_{i1}^G)^2}{\left\|D_{NT}^{-\frac{1}{2}}w_{21}^G\right\|} & 0 & 0 & \frac{-\frac{1}{T}\sum_{t=1}^{T}(f_{t2}^G)^2}{\left\|D_{NT}^{-\frac{1}{2}}w_{21}^G\right\|} & 0 & 0 \\
0 & \frac{\frac{1}{N}\sum_{i=1}^{N}(\lambda_{i1}^G)^2}{\left\|D_{NT}^{-\frac{1}{2}}w_{31}^G\right\|} & 0 & 0 & \frac{-\frac{1}{T}\sum_{t=1}^{T}(f_{t3}^G)^2}{\left\|D_{NT}^{-\frac{1}{2}}w_{31}^G\right\|} & 0 \\
0 & 0 & \frac{\frac{1}{N}\sum_{i=1}^{N}(\lambda_{i2}^G)^2}{\left\|D_{NT}^{-\frac{1}{2}}w_{32}^G\right\|} & 0 & 0 & \frac{-\frac{1}{T}\sum_{t=1}^{T}(f_{t3}^G)^2}{\left\|D_{NT}^{-\frac{1}{2}}w_{32}^G\right\|}
\end{matrix}.$$

Note that $\frac{1}{T}\sum_{t=1}^{T}(f_{tp}^G)^2 = \frac{1}{N}\sum_{i=1}^{N}(\lambda_{ip}^G)^2$ for $p = 1, 2, 3$. Now consider $(plimU^*)g = 0$ for any vector g. If $plim\frac{1}{N}\sum_{i=1}^{N}(\lambda_{i1}^G)^2 \neq plim\frac{1}{N}\sum_{i=1}^{N}(\lambda_{i2}^G)^2$, then $g_1 = g_4 = 0$. If

$plim\frac{1}{N}\sum_{i=1}^{N}(\lambda_{i1}^{G})^2 \neq plim\frac{1}{N}\sum_{i=1}^{N}(\lambda_{i3}^{G})^2$, then $g_2 = g_5 = 0$. And if $plim\frac{1}{N}\sum_{i=1}^{N}(\lambda_{i2}^{G})^2 \neq plim\frac{1}{N}\sum_{i=1}^{N}(\lambda_{i3}^{G})^2$, then $g_3 = g_6 = 0$. Thus by Assumption 6, $g = 0$. ■

Lemma 3 *Under Assumptions 1, 2 and 6, for any $\mathcal{D} > 0$, there exists $C > 0$ and $m > 0$ such that as $(N, T) \to \infty$, $P(\min_{B(\mathcal{D})\cap\left\|D_{NT}^{-\frac{1}{2}}(\phi-\phi^G)\right\|\leq m} \rho_{\min}(-D_{TN}^{-\frac{1}{2}}H(\phi)D_{TN}^{-\frac{1}{2}}) \geq C) \to 1$.*

Proof. The proof is similar to proof of Lemma 2 with some modifications. The intuition behind this lemma is that given we have proved Lemma 2, a small perturbation of ϕ will not affect the order of the largest eigenvalue.

First note that when f and λ are plugged in, the second term on the right hand side of (20), (21) and (22) are no longer zeros. For any vector a with $\|a\| = 1$,

$$\begin{aligned}
& a'(-D_{TN}^{-\frac{1}{2}}H(\phi)D_{TN}^{-\frac{1}{2}})a \\
= \; & a'(-D_{TN}^{-\frac{1}{2}}H_L(\phi)D_{TN}^{-\frac{1}{2}} + c\sum\nolimits_{p=1}^{r} D_{NT}^{-\frac{1}{2}}v_p v_p' D_{NT}^{-\frac{1}{2}})a \\
& +ca'[\sum\nolimits_{j=1}^{r}\sum\nolimits_{k=j+1}^{r}(\frac{1}{N}u_{jk}u_{jk}' + \frac{1}{T}u_{kj}u_{kj}')]a \\
& -N^{-\frac{1}{2}}T^{-\frac{1}{2}}a'J_L(\phi)a \\
& +a'D_{TN}^{-\frac{1}{2}}(\sum\nolimits_{q=1}^{r}\frac{c}{2}NT(\frac{\sum_{i=1}^{N}\lambda_{iq}^2}{N} - \frac{\sum_{t=1}^{T}f_{tq}^2}{T})D_{NT}^{-1}(I_{N+T}\otimes\iota_q)D_{TN}^{-\frac{1}{2}}a \\
& +a'(\sum\nolimits_{p=1}^{r}\sum\nolimits_{q=p+1}^{r}c(\frac{1}{N}\sum\nolimits_{i=1}^{N}\lambda_{ip}\lambda_{iq})D_1)a \\
& +a'(\sum\nolimits_{p=1}^{r}\sum\nolimits_{q=p+1}^{r}c(\frac{1}{T}\sum\nolimits_{t=1}^{T}f_{tp}f_{tq})D_2)a \\
\equiv \; & K_1 + K_2 - K_3 + K_4 + K_5 + K_6.
\end{aligned}$$

Also, using Cauchy-Schwarz inequality, it is easy to show that

$$\max_{\left\|D_{NT}^{-\frac{1}{2}}(\phi-\phi^G)\right\|\leq m}\left|\frac{1}{N}\sum\nolimits_{i=1}^{N}\lambda_{iq}^2 - \frac{1}{N}\sum\nolimits_{i=1}^{N}(\lambda_{iq}^G)^2\right| \leq 2mN^{-\frac{1}{2}}\left\|\lambda^G\right\| + m^2 \text{ for any } q, \quad (33)$$

$$\max_{\left\|D_{NT}^{-\frac{1}{2}}(\phi-\phi^G)\right\|\leq m}\left|\frac{1}{N}\sum\nolimits_{i=1}^{N}\lambda_{ip}\lambda_{iq}\right| \leq 2mN^{-\frac{1}{2}}\left\|\lambda^G\right\| + m^2 \text{ for any } p \neq q, \quad (34)$$

$$\max_{\left\|D_{NT}^{-\frac{1}{2}}(\phi-\phi^G)\right\|\leq m}\left|\frac{1}{T}\sum\nolimits_{t=1}^{T}f_{tq}^2-\frac{1}{T}\sum\nolimits_{t=1}^{T}(f_{tq}^G)^2\right|\leq 2mT^{-\frac{1}{2}}\left\|f^G\right\|+m^2 \text{ for any } q, \quad (35)$$

$$\max_{\left\|D_{NT}^{-\frac{1}{2}}(\phi-\phi^G)\right\|\leq m}\left|\frac{1}{T}\sum\nolimits_{t=1}^{T}f_{tp}f_{tq}\right|\leq 2mT^{-\frac{1}{2}}\left\|f^G\right\|+m^2 \text{ for any } p\neq q. \quad (36)$$

Now we evaluate $K_1,...,K_6$.

(1) Within the neighborhood $B(\mathcal{D})$, $\pi_{it}=f_t'\lambda_i$ is bounded, thus $|\partial_{\pi^2}l_{it}(\pi_{it})|$ is bounded away from zero uniformly. Then similar to the counterpart in the proof of Lemma 2, the smallest nonzero eigenvalue of $-D_{TN}^{-\frac{1}{2}}H_L(\phi)D_{TN}^{-\frac{1}{2}}+c\sum_{p=1}^{r}D_{NT}^{-\frac{1}{2}}v_pv_p'D_{NT}^{-\frac{1}{2}}$ is not smaller than $\rho_{\min}(\frac{c}{N}\sum_{i=1}^{N}\lambda_i\lambda_i')$ and $\rho_{\min}(\frac{c}{T}\sum_{t=1}^{T}f_tf_t')$. By Weyl's inequality, $\rho_{\min}(\frac{c}{N}\sum_{i=1}^{N}\lambda_i\lambda_i')-\rho_{\min}(\frac{c}{N}\sum_{i=1}^{N}\lambda_i^G\lambda_i^{G\prime})$ lies between the smallest and the largest eigenvalues of $\frac{c}{N}\sum_{i=1}^{N}\lambda_i\lambda_i'-\frac{c}{N}\sum_{i=1}^{N}\lambda_i^G\lambda_i^{G\prime}$. This together with (33) and (34) implies that

$$\begin{aligned}
&\max_{\left\|D_{NT}^{-\frac{1}{2}}(\phi-\phi^G)\right\|\leq m}\left|\rho_{\min}(\frac{c}{N}\sum\nolimits_{i=1}^{N}\lambda_i\lambda_i')-\rho_{\min}(\frac{c}{N}\sum\nolimits_{i=1}^{N}\lambda_i^G\lambda_i^{G\prime})\right| \\
\leq\ &\max_{\left\|D_{NT}^{-\frac{1}{2}}(\phi-\phi^G)\right\|\leq m}\left\|\frac{c}{N}\sum\nolimits_{i=1}^{N}\lambda_i\lambda_i'-\frac{c}{N}\sum\nolimits_{i=1}^{N}\lambda_i^G\lambda_i^{G\prime}\right\|_F \\
\leq\ & cr(2mN^{-\frac{1}{2}}\left\|\lambda^G\right\|+m^2).
\end{aligned}$$

It follows that $\min_{\left\|D_{NT}^{-\frac{1}{2}}(\phi-\phi^G)\right\|\leq m}\rho_{\min}(\frac{c}{N}\sum_{i=1}^{N}\lambda_i\lambda_i')$ is not smaller than $\rho_{\min}(\frac{c}{N}\sum_{i=1}^{N}\lambda_i^G\lambda_i^{G\prime})-cr(2mN^{-\frac{1}{2}}\left\|\lambda^G\right\|+m^2)$. Similarly, using Weyl's inequality, (35) and (36), we can show that $\min_{\left\|D_{NT}^{-\frac{1}{2}}(\phi-\phi^G)\right\|\leq m}\rho_{\min}(\frac{c}{T}\sum_{t=1}^{T}f_tf_t')$ is not smaller than $\rho_{\min}(\frac{c}{T}\sum_{t=1}^{T}f_t^Gf_t^{G\prime})-cr(2mT^{-\frac{1}{2}}\left\|f^G\right\|+m^2)$. Note that $N^{-\frac{1}{2}}\left\|\lambda^G\right\|=T^{-\frac{1}{2}}\left\|f^G\right\|=[tr(\frac{1}{T}\sum_{t=1}^{T}f_t^Gf_t^{G\prime})]^{\frac{1}{2}}=tr(\mathcal{V}^{\frac{1}{4}})$. Thus if m is small enough, then there exists $b>0$ such that w.p.a.1, $\min_{\left\|D_{NT}^{-\frac{1}{2}}(\phi-\phi^G)\right\|\leq m}\rho_{\min}(\frac{c}{N}\sum_{i=1}^{N}\lambda_i\lambda_i')\geq b$ and $\min_{\left\|D_{NT}^{-\frac{1}{2}}(\phi-\phi^G)\right\|\leq m}\rho_{\min}(\frac{c}{T}\sum_{t=1}^{T}f_tf_t')\geq b$. It follows that $\min_{B(\mathcal{D})\cap\left\|D_{NT}^{-\frac{1}{2}}(\phi-\phi^G)\right\|\leq m}K_1\geq b\sum_{l=1}^{(N+T)r-r(r-1)}a_l^2$ w.p.a.1.

(2) Similar to the counterpart in the proof of Lemma 2, for some $0 < c_1 < c$, K_2 is not smaller than

$$\begin{aligned}&c_1\rho_{\min}(U^*(\phi)U^*(\phi)')\sum\nolimits_{p=1}^{r}\sum\nolimits_{q=p+1}^{r}(a_{pq}^2 + a_{qp}^2)\\ &-2c_1(\sum\nolimits_{l=1}^{(N+T)r-r(r-1)}a_l^2)^{\frac{1}{2}}\sum\nolimits_{j=1}^{r}\sum\nolimits_{k=j+1}^{r}(\frac{1}{N}\left\|u_{jk}\right\|^2 + \frac{1}{T}\left\|u_{kj}\right\|^2),\end{aligned}$$

where $U^*(\phi)$ is defined the same as U^*, with u_{jk}^G replaced by u_{jk}. Similar to part (1), using Weyl's inequality, (33) and (35), we can show there exists some $M > 0$ such that $\min_{\left\|D_{NT}^{-\frac{1}{2}}(\phi-\phi^G)\right\|\leq m}\rho_{\min}(U^*(\phi)U^*(\phi)')$ is not smaller than $\rho_{\min}(U^*U^{*\prime}) - Mm$ w.p.a.1. Take m small enough, then there exists $d > 0$ such that $\min_{\left\|D_{NT}^{-\frac{1}{2}}(\phi-\phi^G)\right\|\leq m}\rho_{\min}(U^*(\phi)U^*(\phi)') \geq d$ w.p.a.1. Next, since $N^{-\frac{1}{2}}\left\|\lambda^G\right\|$ and $T^{-\frac{1}{2}}\left\|f^G\right\|$ are bounded, both $\max_{\left\|D_{NT}^{-\frac{1}{2}}(\phi-\phi^G)\right\|\leq m}\frac{1}{N}\left\|u_{jk}\right\|^2$ and $\max_{\left\|D_{NT}^{-\frac{1}{2}}(\phi-\phi^G)\right\|\leq m}\frac{1}{T}\left\|u_{kj}\right\|^2$ are bounded by some large M. Thus $\min_{\left\|D_{NT}^{-\frac{1}{2}}(\phi-\phi^G)\right\|\leq m} K_2$ is not smaller than $c_1 d\sum_{p=1}^{r}\sum_{q=p+1}^{r}(a_{pq}^2 + a_{qp}^2) - 2c_1 r(r-1)M(\sum_{l=1}^{(N+T)r-r(r-1)}a_l^2)^{\frac{1}{2}}$ w.p.a.1.

(3) Since $|\partial_{\pi^2}l_{it}(\pi_{it})| \leq b_U$ within $B(\mathcal{D})$ and $\left\|\partial_\pi l(\phi) - \partial_\pi l\right\| \leq \left\|\partial_\pi l(\phi) - \partial_\pi l\right\|_F$, we have

$$\left\|\partial_\pi l(\phi)\right\| \leq \left\|\partial_\pi l\right\| + b_U[\sum\nolimits_{i=1}^{N}\sum\nolimits_{t=1}^{T}(\pi_{it} - \pi_{it}^0)^2]^{\frac{1}{2}}.$$

$\sum_{i=1}^{N}\sum_{t=1}^{T}(\pi_{it}-\pi_{it}^0)^2$ is not larger than the sum of $3\left\|f - f^G\right\|^2\left\|\lambda^G\right\|^2$, $3\left\|f^G\right\|^2\left\|\lambda - \lambda^G\right\|^2$ and $3\left\|f - f^G\right\|^2\left\|\lambda - \lambda^G\right\|^2$, thus $\max_{\left\|D_{NT}^{-\frac{1}{2}}(\phi-\phi^G)\right\|\leq m}\sum_{i=1}^{N}\sum_{t=1}^{T}(\pi_{it} - \pi_{it}^0)^2$ is not larger than $3(m^2T\left\|\lambda^G\right\|^2 + m^2N\left\|f^G\right\|^2 + m^4NT)$. Thus under Assumption 1, there exists $M > 0$ such that $\max_{\left\|D_{NT}^{-\frac{1}{2}}(\phi-\phi^G)\right\|\leq m}[\sum_{i=1}^{N}\sum_{t=1}^{T}(\pi_{it} - \pi_{it}^0)^2]^{\frac{1}{2}} \leq MmN^{\frac{1}{2}}T^{\frac{1}{2}}$ w.p.a.1. Since $|K_3| \leq 2N^{-\frac{1}{2}}T^{-\frac{1}{2}}\left\|J_{L\lambda f'}(\phi)\right\| \leq 2N^{-\frac{1}{2}}T^{-\frac{1}{2}}\left\|\partial_\pi l(\phi)\right\|$, we have $\max_{B(\mathcal{D})\cap\left\|D_{NT}^{-\frac{1}{2}}(\phi-\phi^G)\right\|\leq m}|K_3| \leq 2N^{-\frac{1}{2}}T^{-\frac{1}{2}}\left\|\partial_\pi l\right\| + 2Mm$ w.p.a.1.

(4) First note that $D_{TN}^{-\frac{1}{2}}NTD_{NT}^{-1}(I_{N+T}\otimes\iota_q)D_{TN}^{-\frac{1}{2}} = (I_{N+T}\otimes\iota_q)$. Using(33), (35) and $a'(I_{N+T}\otimes\iota_q)a \leq \left\|a\right\|^2 = 1$ for any q, after some calculation, we have

$$\max_{\left\| D_{NT}^{-\frac{1}{2}}(\phi-\phi^G)\right\| \leq m} |K_4| \leq cr(mN^{-\frac{1}{2}} \left\|\lambda^G\right\| + mT^{-\frac{1}{2}} \left\|f^G\right\| + m^2).$$

(5) Using (34) and $a'D_1a \leq 2\left\|a\right\|^2 = 2$ for any $p \neq q$, after some calculation, we have $\max_{\left\| D_{NT}^{-\frac{1}{2}}(\phi-\phi^G)\right\| \leq m} |K_5| \leq 2cr(r-1)(2mN^{-\frac{1}{2}} \left\|\lambda^G\right\| + m^2)$.

(6) Using (36) and $a'D_2a \leq 2\left\|a\right\|^2 = 2$ for any $p \neq q$, after some calculation, we have $\max_{\left\| D_{NT}^{-\frac{1}{2}}(\phi-\phi^G)\right\| \leq m} |K_6| \leq 2cr(r-1)(2mT^{-\frac{1}{2}} \left\|f^G\right\| + m^2)$.

By Assumption 1, $\max_{\left\| D_{NT}^{-\frac{1}{2}}(\phi-\phi^G)\right\| \leq m} |K_4|$, $\max_{\left\| D_{NT}^{-\frac{1}{2}}(\phi-\phi^G)\right\| \leq m} |K_5|$ and $\max_{\left\| D_{NT}^{-\frac{1}{2}}(\phi-\phi^G)\right\| \leq m} |K_6|$ are all bounded by Mm w.p.a.1. Finally, using the algebra in expression (32) again and taking m small enough, the lemma is proved. ■

Lemma 4 *Let* $U = (u_{12}^G, ..., u_{1r}^G, u_{23}^G, ..., u_{2r}^G, ..., u_{(r-1)r}^G; u_{21}^G, ..., u_{r1}^G, u_{32}^G, ..., u_{r2}^G, ..., u_{r(r-1)}^G)$, *where* u_{pq}^G *is defined in Appendix A. Also, let* $U = (U_\lambda', U_f')'$, U, U_λ, U_f *are of dimension* $(Nr + Tr) \times r(r-1)$, $Nr \times r(r-1)$, $Tr \times r(r-1)$ *respectively. Let* $H_{\lambda\lambda'}$, $H_{\lambda f'}$, $H_{f\lambda'}$, $H_{ff'}$ *be the upper-left, upper-right, lower-left and lower-right block of* H. *Under Assumptions 1, 2 and 3(i), as* $(N, T) \to \infty$,

(i) $\left\|[(V_\lambda, U_\lambda)]_i\right\|_1 = O_p(1)$, *(ii)* $\left\|(V_\lambda, U_\lambda)'\right\|_1 = O_p(1)$,

(iii) $\left\|(V_\lambda, U_\lambda)\right\| = O_p(N^{\frac{1}{2}})$, *(iv)* $\left\|(V_\lambda, U_\lambda)\right\|_1 = O_p(N)$,

(v) $\left\|[H_{L\lambda\lambda'}^{-1}]_i\right\|_1 = O_p(\frac{1}{T})$, *(vi)* $\left\|H_{L\lambda\lambda'}^{-1}\right\|_1 = O_p(\frac{1}{T})$, *(vii)* $\left\|H_{L\lambda\lambda'}^{-1}\right\| = O_p(\frac{1}{T})$,

(viii) $\left\|H_{\lambda\lambda'}^{-1}\right\|_1 = O_p(\frac{1}{T})$, *(ix)* $\left\|H_{\lambda\lambda'}^{-1}\right\| = O_p(\frac{1}{T})$,

(x) $\left\|H_{ff'}^{-1}\right\|_1 = O_p(\frac{1}{N})$, *(xi)* $\left\|H_{ff'}^{-1}\right\| = O_p(\frac{1}{N})$,

(xii) $\max_i \left\|[H_{\lambda f'}]_i\right\| = O_p(N^{\frac{1}{\xi}}T^{\frac{1}{2}+\frac{1}{\xi}})$, *(xiii)* $\left\|[H_{\lambda f'}]_i\right\| = O_p(T^{\frac{1}{2}})$,

(xiv) $\left\|[H_{\lambda f'}]_i\right\|_1 = O_p(T^{\frac{1}{\xi}})$, *(xv)* $\left\|H_{\lambda f'}\right\|_1 = O_p(N^{1+\frac{1}{\xi}}T^{\frac{1}{\xi}})$, *(xvi)* $\left\|H_{\lambda f'}\right\| = O_p(N^{\frac{1}{2}}T^{\frac{1}{2}})$.

Proof. Parts (i)-(iv): Obvious.

Parts (v)-(vii): Noting that $H_{L\lambda\lambda'}$ is block diagonal, we have

$$\begin{aligned} \left\|H_{L\lambda\lambda'}^{-1}\right\|_1 &= \max_i \left\|(\sum_{t=1}^T \partial_{\pi^2} l_{it} f_t^0 f_t^{0\prime})^{-1}\right\|_1 \leq \max_i \sqrt{r} \left\|(\sum_{t=1}^T \partial_{\pi^2} l_{it} f_t^0 f_t^{0\prime})^{-1}\right\| \\ &\leq \frac{\sqrt{r}}{b_L} \left\|(\sum_{t=1}^T f_t^0 f_t^{0\prime})^{-1}\right\| = O_p(\frac{1}{T}). \end{aligned} \tag{37}$$

From expression (37), we can see that $\left\|[H_{L\lambda\lambda'}^{-1}]_i\right\|_1$ and $\left\|H_{L\lambda\lambda'}^{-1}\right\|$ are also $O_p(\frac{1}{T})$.

Parts (viii)-(ix): By equations (19) and (23), $H_{\lambda\lambda'} = H_{L\lambda\lambda'} - c\frac{T}{N}(V_\lambda, U_\lambda)(V_\lambda, U_\lambda)'$. Thus by Woodbury identity,

$$H_{\lambda\lambda'}^{-1} = H_{L\lambda\lambda'}^{-1} - H_{L\lambda\lambda'}^{-1}(V_\lambda, U_\lambda)[-\frac{N}{cT}I_{r^2} + (V_\lambda, U_\lambda)'H_{L\lambda\lambda'}^{-1}(V_\lambda, U_\lambda)]^{-1}(V_\lambda, U_\lambda)'H_{L\lambda\lambda'}^{-1}. \quad (38)$$

By positive definiteness of $-(V_\lambda, U_\lambda)'H_{L\lambda\lambda'}^{-1}(V_\lambda, U_\lambda)$, we have

$$\begin{aligned} & \left\|[-\frac{N}{cT}I_{r^2} + (V_\lambda, U_\lambda)'H_{L\lambda\lambda'}^{-1}(V_\lambda, U_\lambda)]^{-1}\right\|_1 \\ \leq \; & r\left\|[-\frac{N}{cT}I_{r^2} + (V_\lambda, U_\lambda)'H_{L\lambda\lambda'}^{-1}(V_\lambda, U_\lambda)]^{-1}\right\| \leq rc\frac{T}{N}. \end{aligned} \quad (39)$$

Part (viii) follows from part (ii), part (iv), part (vi) and expression (39). Part (ix) follows from part (iii), part (vii) and expression (39).

Parts (x)-(xi): The proof is similar to parts (viii)-(ix).

Parts (xii)-(xvi): First note that $H_{\lambda f'} = H_{L\lambda f'} + H_{P\lambda f'} + J_{L\lambda f'}$.

For (xii) and (xiii), we have $\|[H_{L\lambda f'}]_i\| \leq b_U \left\|\lambda_i^G\right\| \left\|f^G\right\|$, $\|[H_{P\lambda f'}]_i\| \leq c\left\|\lambda_i^G\right\| \left\|f^G\right\|$ and $\|[J_{L\lambda f'}]_i\| \leq (r\sum_{t=1}^T(\partial_\pi l_{it})^2)^{\frac{1}{2}}$.

For (xiv), we have $\|[J_{L\lambda f'}]_i\|_1 \leq \max_t |\partial_\pi l_{it}|$, $\|[H_{L\lambda f'}]_i\|_1 \leq rb_U \left\|\lambda_i^G\right\| \max_t \left\|f_t^G\right\|$ and $\|[H_{P\lambda f'}]_i\|_1 \leq cr\left\|\lambda_i^G\right\| \max_t \left\|f_t^G\right\|$.

For (xv), we have $\|J_{L\lambda f'}\|_1 \leq \max_t \sum_{i=1}^N |\partial_\pi l_{it}|$, $\|H_{L\lambda f'}\|_1 \leq rb_U \sum_{i=1}^N \left\|\lambda_i^G\right\| \max_t \left\|f_t^G\right\|$ and $\|H_{P\lambda f'}\|_1 \leq cr\sum_{i=1}^N \left\|\lambda_i^G\right\| \max_t \left\|f_t^G\right\|$.

For (xvi), we have $\|H_{P\lambda f'}\| \leq c\left\|\lambda^G\right\| \left\|f^G\right\|$, $\|H_{L\lambda f'}\| \leq b_U \left\|\lambda^G\right\| \left\|f^G\right\|$ and by Lemma 1, $\|J_{L\lambda f'}\| \leq \|\partial_\pi l\| = O_p(N^{\frac{1}{2}}T^{\frac{1}{4}} + N^{\frac{1}{4}}T^{\frac{1}{2}})$.

Also note that by Assumption 3(i), $\max_t |\partial_\pi l_{it}|$ is $O_p(T^{\frac{1}{\xi}})$, $\max_i |\partial_\pi l_{it}|$ is $O_p(N^{\frac{1}{\xi}})$ and $\max_{i,t} |\partial_\pi l_{it}|$ is $O_p(N^{\frac{1}{\xi}}T^{\frac{1}{\xi}})$. ■

Lemma 5 *Under Assumptions 1, 2, 3(i) and 6, as $(N,T) \to \infty$, the 1-norm of the upper-left, lower-right, upper-right and lower-left block of H^{-1} is $O_p(\frac{N^{\frac{2}{\xi}}T^{\frac{2}{\xi}}}{T})$, $O_p(\frac{N^{\frac{2}{\xi}}T^{\frac{2}{\xi}}}{N})$, $O_p(\frac{N^{\frac{3}{\xi}}T^{\frac{3}{\xi}}}{T})$ and $O_p(\frac{N^{\frac{3}{\xi}}T^{\frac{3}{\xi}}}{N})$ respectively.*

Proof. (1) The upper-left block of H^{-1} is:

$$[H_{\lambda\lambda'} - H_{\lambda f'}H_{ff'}^{-1}H_{f\lambda'}]^{-1} = H_{\lambda\lambda'}^{-1} + H_{\lambda\lambda'}^{-1}H_{\lambda f'}[H_{ff'} - H_{f\lambda'}H_{\lambda\lambda'}^{-1}H_{\lambda f'}]^{-1}H_{f\lambda'}H_{\lambda\lambda'}^{-1}. \quad (40)$$

From Lemma 4(viii), we have $\left\|H_{\lambda\lambda'}^{-1}\right\|_1 = O_p(\frac{1}{T})$. We next show

$$\left\|H_{\lambda f'}[H_{ff'} - H_{f\lambda'}H_{\lambda\lambda'}^{-1}H_{\lambda f'}]^{-1}H_{f\lambda'}\right\|_1 = O_p(N^{\frac{2}{\xi}}T^{1+\frac{2}{\xi}}). \tag{41}$$

Let $\|A\|_{\max}$ be the max norm of matrix A. It suffices to show

$$\left\|H_{\lambda f'}[H_{ff'} - H_{f\lambda'}H_{\lambda\lambda'}^{-1}H_{\lambda f'}]^{-1}H_{f\lambda'}\right\|_{\max} = O_p(\frac{T}{N}N^{\frac{2}{\xi}}T^{\frac{2}{\xi}}). \tag{42}$$

The (ip, jq) element is $[H_{\lambda f'}]_{ip}[H_{ff'}-H_{f\lambda'}H_{\lambda\lambda'}^{-1}H_{\lambda f'}]^{-1}[H_{\lambda f'}]'_{jq}$. $[H_{ff'}-H_{f\lambda'}H_{\lambda\lambda'}^{-1}H_{\lambda f'}]^{-1}$ equals the lower right block of $D_{TN}^{-\frac{1}{2}}(D_{TN}^{-\frac{1}{2}}HD_{TN}^{-\frac{1}{2}})^{-1}D_{TN}^{-\frac{1}{2}}$, thus by Lemma 2,

$$\left\|[H_{ff'} - H_{f\lambda'}H_{\lambda\lambda'}^{-1}H_{\lambda f'}]^{-1}\right\| \leq \frac{1}{N}\left\|(D_{TN}^{-\frac{1}{2}}HD_{TN}^{-\frac{1}{2}})^{-1}\right\| = O_p(\frac{1}{N}). \tag{43}$$

This together with Lemma 4(xii) proves (42). Thus the 1-norm of the upper-left block is $O_p(\frac{N^{\frac{2}{\xi}}T^{\frac{2}{\xi}}}{T})$.

(2) The lower-right block is $[H_{ff'} - H_{f\lambda'}H_{\lambda\lambda'}^{-1}H_{\lambda f'}]^{-1}$, thus by symmetry, its magnitude is $O_p(\frac{N^{\frac{2}{\xi}}T^{\frac{2}{\xi}}}{N})$.

(3) The upper-right block is $-[H_{\lambda\lambda'} - H_{\lambda f'}H_{ff'}^{-1}H_{f\lambda'}]^{-1}H_{\lambda f'}H_{ff'}^{-1}$. Part (1), parts (x) and (xv) of Lemma 4 together implies this term is $O_p(\frac{N^{\frac{3}{\xi}}T^{\frac{3}{\xi}}}{T})$.

(4) The lower-left block is the transpose of the upper-right block and $\left\|H_{f\lambda'}\right\|_1 = O_p(N^{\frac{1}{\xi}}T^{1+\frac{1}{\xi}})$, thus is $O_p(\frac{N^{\frac{3}{\xi}}T^{\frac{3}{\xi}}}{N})$. ∎

Lemma 6 *Under Assumptions 1, 2, 4 and 6, as* $(N, T) \to \infty$, $\left\|\tilde{H}_{\lambda\lambda'}^{-1}\right\|_1 = O_p(\frac{1}{T})$ *and* $\left\|\tilde{H}_{ff'}^{-1}\right\|_1 = O_p(\frac{1}{N})$.

Proof. For V and U defined in the proof of Lemma 2, when $\phi^G + s(\hat{\phi}^* - \phi^G)$ is plugged in, use notation $V(s)$, $V_\lambda(s)$, $V_f(s)$, $U(s)$, $U_\lambda(s)$ and $U_f(s)$. It follows that $V_\lambda(s) = V_\lambda(0) + s(V_\lambda(1) - V_\lambda(0))$ and

$$\begin{aligned}\tilde{H}_{\lambda\lambda'} &= \int_0^1 H_{\lambda\lambda'}(s)ds = \int_0^1 H_{L\lambda\lambda'}(s)ds + \int_0^1 H_{P\lambda\lambda'}(s)ds \\ &= \tilde{H}_{L\lambda\lambda'} - c\frac{T}{N}\int_0^1 (V_\lambda(s), U_\lambda(s))(V_\lambda(s), U_\lambda(s))'ds.\end{aligned}$$

Since $\int_0^1 sds = 1/2$ and $\int_0^1 s^2 ds = 1/3$, we have

$$\begin{aligned}\int_0^1 V_\lambda(s)V_\lambda(s)'ds &= V_\lambda(0)V_\lambda(0)' + \frac{1}{3}(V_\lambda(1) - V_\lambda(0))(V_\lambda(1) - V_\lambda(0))' \\ &\quad + \frac{1}{2}V_\lambda(0)(V_\lambda(1) - V_\lambda(0))' + \frac{1}{2}(V_\lambda(1) - V_\lambda(0))V_\lambda(0)' \\ &= V_\lambda(\frac{1}{2})V_\lambda(\frac{1}{2})' + \frac{1}{12}(V_\lambda(1) - V_\lambda(0))(V_\lambda(1) - V_\lambda(0))'.\end{aligned}$$

Similarly, we also have

$$\int_0^1 U_\lambda(s)U_\lambda(s)'ds = U_\lambda(\frac{1}{2})U_\lambda(\frac{1}{2})' + \frac{1}{12}(U_\lambda(1) - U_\lambda(0))(U_\lambda(1) - U_\lambda(0))'.$$

It follows that $\tilde{H}_{\lambda\lambda'} = \tilde{H}_{L\lambda\lambda'} - c\frac{T}{N}BB'$, where

$$B \equiv (V_\lambda(\frac{1}{2}), U_\lambda(\frac{1}{2}), (V_\lambda(1) - V_\lambda(0))/2\sqrt{3}, (U_\lambda(1) - U_\lambda(0))/2\sqrt{3}). \tag{44}$$

Thus by Woodbury identity,

$$\tilde{H}^{-1}_{\lambda\lambda'} = \tilde{H}^{-1}_{L\lambda\lambda'} + \tilde{H}^{-1}_{L\lambda\lambda'}B[\frac{N}{cT}I_{2r^2} - B'\tilde{H}^{-1}_{L\lambda\lambda'}B]^{-1}B'\tilde{H}^{-1}_{L\lambda\lambda'}.$$

Consider $\tilde{H}^{-1}_{L\lambda\lambda'}$ first. $\tilde{H}_{L\lambda\lambda'}$ is block-diagonal with $\tilde{H}_{L\lambda_i\lambda_i'}$ as the i-th block. Thus $\tilde{H}^{-1}_{L\lambda\lambda'}$ is also block-diagonal and the i-th block is $\tilde{H}^{-1}_{L\lambda_i\lambda_i'}$. It follows that

$$\left\|\tilde{H}^{-1}_{L\lambda\lambda'}\right\|_1 = \max_i \left\|\tilde{H}^{-1}_{L\lambda_i\lambda_i'}\right\|_1 \leq \max_i r^{\frac{1}{2}}\left\|\tilde{H}^{-1}_{L\lambda_i\lambda_i'}\right\|.$$

Due to the four facts listed below, $\min_i \rho_{\min}(-\tilde{H}_{L\lambda_i\lambda_i'}) \geq TCb_L/2$ w.p.a.1. This implies $\max_i \left\|\tilde{H}^{-1}_{L\lambda_i\lambda_i'}\right\| \leq 2/TCb_L$ w.p.a.1, thus $\left\|\tilde{H}^{-1}_{L\lambda\lambda'}\right\|_1$ is $O_p(\frac{1}{T})$.

1. $\rho_{\min}(-\tilde{H}_{L\lambda_i\lambda_i'}) \geq \int_0^1 \rho_{\min}(-H_{L\lambda_i\lambda_i'}(s))ds \geq \min_{0\leq s\leq 1}\rho_{\min}(-H_{L\lambda_i\lambda_i'}(s))$, where the first inequality follows from continuity of the smallest eigenvalues and Weyl's inequality.

2. $\rho_{\min}(-H_{L\lambda_i\lambda_i'}(s)) \geq b_L\rho_{\min}(\sum_{t=1}^T f_t^G f_t^{G\prime}) - 2b_L s(\sum_{t=1}^T \left\|f_t^G\right\|^2 \sum_{t=1}^T \left\|\hat{f}_t^* - f_t^G\right\|^2)^{\frac{1}{2}} - b_L s^2\sum_{t=1}^T \left\|\hat{f}_t^* - f_t^G\right\|^2$ for any i, because $\rho_{\min}(A) \geq \rho_{\min}(B) - \|A - B\|_F$ for symmetric matrices A and B, and $-\partial_{\pi^2} l_{it}(\pi_{it})$ is uniformly bounded below by b_L

within the neighborhood $B(\mathcal{D})$ and $\hat{\phi}^*$ lies in $B(\mathcal{D})$.

3. By Assumption 1, there exists some $C > 0$ such that $\rho_{\min}(\sum_{t=1}^{T} f_t^G f_t^{G\prime}) \geq TC$ w.p.a.1, and there exists some $M > 0$ such that $\sum_{t=1}^{T} \left\| f_t^G \right\|^2 \leq MT$.

4. Because $\hat{\phi}^*$ lies in $B(\mathcal{D}) \cap \left\| D_{NT}^{-\frac{1}{2}}(\phi - \phi^G) \right\| \leq m$, $\sum_{t=1}^{T} \left\| \hat{f}_t^* - f_t^G \right\|^2 \leq m^2 T$. Take m small enough.

Next, by expression (44) and the fact that both λ^G and $\hat{\lambda}^*$ lie in $B(\mathcal{D})$, we have

$$\begin{aligned} \|B\|_1 &\leq \left\|\lambda^G\right\|_1 + \left\|\hat{\lambda}^* - \lambda^G\right\|_1 = O(N), \\ \|B'\|_1 &= \|B\|_\infty \leq \left\|\lambda^G\right\|_\infty + \left\|\hat{\lambda}^* - \lambda^G\right\|_\infty = O(1). \end{aligned}$$

When $-\tilde{H}_{L\lambda\lambda'}^{-1}$ is positive definite, $\rho_{\min}(\frac{N}{cT} I_{2r^2} - B'\tilde{H}_{L\lambda\lambda'}^{-1} B)$ is not smaller than $\frac{N}{cT}$ and $\left\| [\frac{N}{cT} I_{2r^2} - B'\tilde{H}_{L\lambda\lambda'}^{-1} B]^{-1} \right\|_1$ is not larger than $\sqrt{2r^2} \left\| [\frac{N}{cT} I_{2r^2} - B'\tilde{H}_{L\lambda\lambda'}^{-1} B]^{-1} \right\| \leq \sqrt{2r^2} \frac{cT}{N}$. Since $-\tilde{H}_{L\lambda\lambda'}^{-1}$ is positive definite w.p.a.1, $\left\| [\frac{N}{cT} I_{2r^2} - B'\tilde{H}_{L\lambda\lambda'}^{-1} B]^{-1} \right\|_1 \leq \sqrt{2r^2} \frac{cT}{N}$ also holds w.p.a.1.

Taking all above together, $\left\| \tilde{H}_{\lambda\lambda'}^{-1} \right\|_1 = O_p(\frac{1}{T})$. By symmetry, $\left\| \tilde{H}_{ff'}^{-1} \right\|_1 = O_p(\frac{1}{N})$. ■

Lemma 7 *Under Assumptions 1, 2, 3(i), 4 and 6, as $(N, T) \to \infty$, the 1-norm of the upper-left, lower-right, upper-right and lower-left block of $\tilde{H}^{-1}$ is $O_p(\frac{N^{\frac{2}{\xi}} T^{\frac{2}{\xi}}}{T})$, $O_p(\frac{N^{\frac{2}{\xi}} T^{\frac{2}{\xi}}}{N})$, $O_p(\frac{N^{\frac{3}{\xi}} T^{\frac{3}{\xi}}}{T})$ and $O_p(\frac{N^{\frac{3}{\xi}} T^{\frac{3}{\xi}}}{N})$ respectively.*

Proof. Based on the following facts, the proof is similar to the proof of Lemma 5.

(1) $\left\| \tilde{H}_{\lambda\lambda'}^{-1} \right\|_1 = O_p(\frac{1}{T})$.

(2) $\left\| \tilde{H}_{ff'}^{-1} \right\|_1 = O_p(\frac{1}{N})$.

(3) $\left\| [\tilde{H}_{ff'} - \tilde{H}_{f\lambda'} \tilde{H}_{\lambda\lambda'}^{-1} \tilde{H}_{\lambda f'}]^{-1} \right\| = O_p(\frac{1}{N})$.

(4) $\max_i \left\| [\tilde{H}_{\lambda f'}]_i \right\| = O_p(N^{\frac{1}{\xi}} T^{\frac{1}{2} + \frac{1}{\xi}})$.

(5) $\left\| \tilde{H}_{\lambda f'} \right\|_1 = O_p(N^{1+\frac{1}{\xi}} T^{\frac{1}{\xi}})$.

(1) and (2) follow from Lemma 6. For (3), $\left\| [\tilde{H}_{ff'} - \tilde{H}_{f\lambda'} \tilde{H}_{\lambda\lambda'}^{-1} \tilde{H}_{\lambda f'}]^{-1} \right\|$ is not larger than $\frac{1}{N} \left\| (D_{TN}^{-\frac{1}{2}} \tilde{H} D_{TN}^{-\frac{1}{2}})^{-1} \right\|$ because $[\tilde{H}_{ff'} - \tilde{H}_{f\lambda'} \tilde{H}_{\lambda\lambda'}^{-1} \tilde{H}_{\lambda f'}]^{-1}$ is the lower-right block of $\tilde{H}^{-1}$, which equals $\frac{1}{N}$ times the lower right block of $(D_{TN}^{-\frac{1}{2}} \tilde{H} D_{TN}^{-\frac{1}{2}})^{-1}$. Due

to continuity of the smallest eigenvalue and Weyl's inequality, $\rho_{\min}(-D_{TN}^{-\frac{1}{2}}\tilde{H}D_{TN}^{-\frac{1}{2}}) \geq \min_{0\leq s\leq 1}\rho_{\min}(-D_{TN}^{-\frac{1}{2}}H(s)D_{TN}^{-\frac{1}{2}})$. This together with Lemma 3 and the fact that $\hat{\phi}^*$ lies in $B(\mathcal{D})\cap\left\|D_{NT}^{-\frac{1}{2}}(\phi-\phi^G)\right\|\leq m$ implies that $\left\|(-D_{TN}^{-\frac{1}{2}}\tilde{H}D_{TN}^{-\frac{1}{2}})^{-1}\right\|$ is $O_p(1)$.

For (4) and (5), note that $\max_i\left\|[\tilde{H}_{\lambda f'}-H_{\lambda f'}]_i\right\|\leq\max_i\max_{0\leq s\leq 1}\|[H_{\lambda f'}(s)-H_{\lambda f'}]_i\|\leq T^{\frac{1}{2}}\max_{0\leq s\leq 1}\|H_{\lambda f'}(s)-H_{\lambda f'}\|_{\max}$ and $\left\|\tilde{H}_{\lambda f'}-H_{\lambda f'}\right\|_1\leq N\max_{0\leq s\leq 1}\|H_{\lambda f'}(s)-H_{\lambda f'}\|_{\max}$. Since $\hat{\phi}^*$ lies in $B(\mathcal{D})\cap\left\|D_{NT}^{-\frac{1}{2}}(\phi-\phi^G)\right\|\leq m$ and $|\partial_{\pi^2}l_{it}(\cdot)|$, $|\partial_{\pi^3}l_{it}(\cdot)|$ and $f_t'\lambda_i$ are all bounded within the neighborhood $B(\mathcal{D})$, $\max_{0\leq s\leq 1}\|H_{\lambda f'}(s)-H_{\lambda f'}\|_{\max}$ is $O_p(1)$. It follows that $\max_i\left\|[\tilde{H}_{\lambda f'}-H_{\lambda f'}]_i\right\|$ is $O_p(T^{\frac{1}{2}})$ and $\left\|\tilde{H}_{\lambda f'}-H_{\lambda f'}\right\|_1$ is $O_p(N)$. These together with parts (xii) and (xv) of Lemma 4 proves (4) and (5). ■

Lemma 8 *Under Assumptions 1-5, as* $(N,T)\to\infty$,

(i) $\left\|(V_\lambda,U_\lambda)'H_{L\lambda\lambda'}^{-1}S_\lambda\right\|=O_p(N^{\frac{1}{2}}T^{-\frac{1}{2}})$, *and* $\left\|(V_f,U_f)'H_{Lff'}^{-1}S_f\right\|=O_p(N^{-\frac{1}{2}}T^{\frac{1}{2}})$,

(ii) $\left\|(H_{Lf\lambda'}+H_{Pf\lambda'})H_{\lambda\lambda'}^{-1}S_\lambda\right\|=O_p(N^{\frac{1}{2}})$, *and* $\left\|(H_{L\lambda f'}+H_{P\lambda f'})H_{ff'}^{-1}S_f\right\|=O_p(T^{\frac{1}{2}})$,

(iii) $\left\|[(H_{Lf\lambda'}+H_{Pf\lambda'})H_{\lambda\lambda'}^{-1}S_\lambda]_s\right\|=O_p(N^{\frac{1}{2}}T^{-\frac{1}{2}})$, *and* $\left\|[(H_{L\lambda f'}+H_{P\lambda f'})H_{ff'}^{-1}S_f]_i\right\|=O_p(N^{-\frac{1}{2}}T^{\frac{1}{2}})$,

(iv) $\left\|[J_{Lf\lambda'}H_{\lambda\lambda'}^{-1}S_\lambda]_s\right\|=O_p(N^{\frac{1}{2}}T^{-\frac{1}{2}})$, *and* $\left\|[J_{L\lambda f'}H_{ff'}^{-1}S_f]_i\right\|=O_p(N^{-\frac{1}{2}}T^{\frac{1}{2}})$,

(v) $\left\|J_{Lf\lambda'}H_{\lambda\lambda'}^{-1}S_\lambda\right\|=O_p(N^{\frac{1}{2}})$, *and* $\left\|J_{L\lambda f'}H_{ff'}^{-1}S_f\right\|=O_p(T^{\frac{1}{2}})$.

Proof. Part (i): It suffices to show the first half. $(V_\lambda,U_\lambda)'H_{L\lambda\lambda'}^{-1}S_\lambda$ is a r^2 dimensional vector. From the definition of V_λ and U_λ, we need to show that for any p and q, $\sum_{i=1}^N\lambda_{ip}^G 1_q^{(r)\prime}(\sum_{t=1}^T\partial_{\pi^2}l_{it}f_t^Gf_t^{G\prime})^{-1}(\sum_{t=1}^T(\partial_\pi l_{it})f_t^G)$ is $O_p(N^{\frac{1}{2}}T^{-\frac{1}{2}})$. Thus it suffices to show that $\left\|\sum_{i=1}^N\sum_{t=1}^T(\sum_{t=1}^T\partial_{\pi^2}l_{it}f_t^Gf_t^{G\prime})^{-1}f_t^G\lambda_i^{G\prime}\partial_\pi l_{it}\right\|_F$ is $O_p(N^{\frac{1}{2}}T^{-\frac{1}{2}})$. This is equivalent to Assumption 5(ii) because $f_t^G=G'f_t^0$, $\lambda_i^G=G^{-1}\lambda_i^0$, and the Frobenius norm and spectral norm are equivalent for fixed dinensional matrices.

Parts (ii) and (iii): It suffices to show the first half. From equation (38) we have

$$[H_{Lf\lambda'}+H_{Pf\lambda'})H_{\lambda\lambda'}^{-1}S_\lambda]_s=[(H_{Lf\lambda'}+H_{Pf\lambda'})H_{L\lambda\lambda'}^{-1}S_\lambda]_s-[H_{Lf\lambda'}+H_{Pf\lambda'}]_sH_{L\lambda\lambda'}^{-1}$$
$$(V_\lambda,U_\lambda)[-\frac{N}{cT}I_{r^2}+(V_\lambda,U_\lambda)'H_{L\lambda\lambda'}^{-1}(V_\lambda,U_\lambda)]^{-1}(V_\lambda,U_\lambda)'H_{L\lambda\lambda'}^{-1}S_\lambda. \tag{45}$$

Consider the first term on the right hand side. Consider $H_{Pf\lambda'}H_{L\lambda\lambda'}^{-1}S_\lambda$ first. The q-th element in the s-th block is $-cf_{sq}^G\sum_{i=1}^N\lambda_{iq}^G1_q^{(r)\prime}(\sum_{t=1}^T(\partial_{\pi^2}l_{it})f_t^Gf_t^{G\prime})^{-1}(\sum_{t=1}^T\partial_\pi l_{it}f_t^G)$.

Part (i) shows that $\sum_{i=1}^{N}\lambda_{iq}^{G}1_q^{(r)\prime}(\sum_{t=1}^{T}(\partial_{\pi^2}l_{it})f_t^G f_t^{G\prime})^{-1}(\sum_{t=1}^{T}\partial_\pi l_{it}f_t^G)$ is $O_p(N^{\frac{1}{2}}T^{-\frac{1}{2}})$. Since $\sum_{s=1}^{T}\left\|f_s^G\right\|^2$ is $O_p(T)$ and $\left\|f_s^G\right\|$ is $O_p(1)$, $\left\|[H_{Pf\lambda'}H_{L\lambda\lambda'}^{-1}S_\lambda]_s\right\|$ is $O_p(N^{\frac{1}{2}}T^{-\frac{1}{2}})$ and $\left\|H_{Pf\lambda'}H_{L\lambda\lambda'}^{-1}S_\lambda\right\|$ is $O_p(N^{\frac{1}{2}})$. Next consider $H_{Lf\lambda'}H_{L\lambda\lambda'}^{-1}S_\lambda$.

$$
\begin{aligned}
G[H_{Lf\lambda'}H_{L\lambda\lambda'}^{-1}S_\lambda]_s &= G\sum_{i=1}^{N}\partial_{\pi^2}l_{is}\lambda_i^G f_s^{G\prime}(\sum_{t=1}^{T}\partial_{\pi^2}l_{it}f_t^G f_t^{G\prime})^{-1}(\sum_{t=1}^{T}\partial_\pi l_{it}f_t^G) \\
&= \sum_{i=1}^{N}\partial_{\pi^2}l_{is}\lambda_i^0 f_s^{0\prime}(\sum_{t=1}^{T}\partial_{\pi^2}l_{it}f_t^0 f_t^{0\prime})^{-1}(\sum_{t=1}^{T}\partial_\pi l_{it}f_t^0) \\
&= [f_s^{0\prime}\sum_{i=1}^{N}\sum_{t=1}^{T}(\sum_{t=1}^{T}\partial_{\pi^2}l_{it}f_t^0 f_t^{0\prime})^{-1}\partial_\pi l_{it}f_t^0\lambda_i^{0\prime}\partial_{\pi^2}l_{is}]'
\end{aligned}
$$

Thus by Assumption 5(ii), $\left\|[H_{Lf\lambda'}H_{L\lambda\lambda'}^{-1}S_\lambda]_s\right\|$ is $O_p(N^{\frac{1}{2}}T^{-\frac{1}{2}})$ and $\left\|H_{Lf\lambda'}H_{L\lambda\lambda'}^{-1}S_\lambda\right\|$ is $O_p(N^{\frac{1}{2}})$.

Now consider the second term on the right hand side of (45). By Assumptions 1 and 2(ii), $\left\|[H_{Lf\lambda'}+H_{Pf\lambda'}]_s\right\| = O_p(N^{\frac{1}{2}})$ and $\left\|H_{Lf\lambda'}+H_{Pf\lambda'}\right\|$ is $O_p(N^{\frac{1}{2}}T^{\frac{1}{2}})$. These together with Lemma 4(iii), Lemma 4(vii), inequality (39) and part (i) finishes the proof.

Parts (iv) and (v): It suffices to show the first half. Similar to expression (45),

$$
\begin{aligned}
&[J_{Lf\lambda'}H_{\lambda\lambda'}^{-1}S_\lambda]_s = [J_{Lf\lambda'}H_{L\lambda\lambda'}^{-1}S_\lambda]_s - [J_{Lf\lambda'}]_s H_{L\lambda\lambda'}^{-1}(V_\lambda, \\
&U_\lambda)[-\frac{N}{cT}I_{r^2} + (V_\lambda,U_\lambda)'H_{L\lambda\lambda'}^{-1}(V_\lambda,U_\lambda)]^{-1}(V_\lambda,U_\lambda)'H_{L\lambda\lambda'}^{-1}S_\lambda. \qquad (46)
\end{aligned}
$$

The second term on the right hand side of (46) is $O_p(N^{\frac{1}{2}}T^{-\frac{1}{2}})$. The proof is the same as the second term on the right hand side of (45) except that here we use $\left\|[J_{Lf\lambda'}]_s\right\| = O_p(N^{\frac{1}{2}})$. Now consider the first term.

$$
G[J_{Lf\lambda'}H_{L\lambda\lambda'}^{-1}S_\lambda]_s = \sum_{i=1}^{N}\sum_{t=1}^{T}(\sum_{t=1}^{T}\partial_{\pi^2}l_{it}f_t^0 f_t^{0\prime})^{-1}\partial_\pi l_{is}\partial_\pi l_{it}f_t^0.
$$

Thus by Assumption 5(i), $\left\|[J_{Lf\lambda'}H_{L\lambda\lambda'}^{-1}S_\lambda]_s\right\|$ is $O_p(N^{\frac{1}{2}}T^{-\frac{1}{2}})$ and $\left\|J_{Lf\lambda'}H_{L\lambda\lambda'}^{-1}S_\lambda\right\|$ is $O_p(N^{\frac{1}{2}})$. ■

Lemma 9 *Following the definitions of R, R_λ, R_f, $R_{\lambda,iq}$, $R_{f,tq}$, ϕ_{iq}^* and ϕ_{tq}^* in Section 4, under Assumptions 1-4, 6 and 7, as $(N,T)\to\infty$,*

$\left\|[R_\lambda]_i\right\|_1 = O_p(\frac{T}{\delta_{NT}^2})$ *for each i, and* $\left\|R_\lambda\right\|_1 = O_p(\frac{NT}{\delta_{NT}^2})$,

$\left\|[R_f]_t\right\|_1 = O_p(\frac{N}{\delta_{NT}^2})$ *for each t, and* $\left\|R_f\right\|_1 = O_p(\frac{NT}{\delta_{NT}^2})$.

Proof. Note that $R_{\lambda,iq} = (\hat{\phi}-\phi^G)'\partial_{\phi\phi'\lambda_{iq}}L(\phi^*_{iq})(\hat{\phi}-\phi^G)+(\hat{\phi}-\phi^G)'\partial_{\phi\phi'\lambda_{iq}}P(\phi^*_{iq})(\hat{\phi}-\phi^G)$.

The first term on the right hand side equals $(\hat{\lambda}_i-\lambda_i^G)'\partial_{\lambda_i\lambda_i'\lambda_{iq}}L(\phi^*_{iq})(\hat{\lambda}_i-\lambda_i^G)+2\sum_{t=1}^T(\hat{\lambda}_i-\lambda_i^G)'\partial_{\lambda_i f_t'\lambda_{iq}}L(\phi^*_{iq})(\hat{f}_t-f_t^G)+\sum_{t=1}^T(\hat{f}_t-f_t^G)'\partial_{f_t f_t'\lambda_{iq}}L(\phi^*_{iq})(\hat{f}_t-f_t^G)\equiv L1i+L2i+L3i$. Based on the expressions of $\partial_{\lambda_i\lambda_i'}L(\phi)$, $\partial_{\lambda_i f_t'}L(\phi)$ and $\partial_{f_t f_t'}L(\phi)$ in equations (16) and (17), it can be verified that

$$
\begin{aligned}
\partial_{\lambda_i\lambda_i'\lambda_{iq}}L(\phi) &= \sum\nolimits_{t=1}^T \partial_{\pi^3}l_{it}(\pi_{it})f_t f_t' f_{tq},\\
\partial_{\lambda_i f_t'\lambda_{iq}}L(\phi) &= \partial_{\pi^3}l_{it}(\pi_{it})f_t\lambda_i' f_{tq}+\partial_{\pi^2}l_{it}(\pi_{it})I_r f_{tq}+\partial_{\pi^2}l_{it}(\pi_{it})f_t 1_q^{(r)\prime},\\
\partial_{f_t f_t'\lambda_{iq}}L(\phi) &= \partial_{\pi^3}l_{it}(\pi_{it})\lambda_i\lambda_i' f_{tq}+\partial_{\pi^2}l_{it}(\pi_{it})1_q^{(r)}\lambda_i'+\partial_{\pi^2}l_{it}(\pi_{it})\lambda_i 1_q^{(r)\prime}.
\end{aligned}
$$

Since $\hat{\phi}$ (consequently ϕ^*_{iq} and ϕ^*_{tq}) lies in $B(\mathcal{D})$ w.p.a.1 and $\|f_t\|$, $\|\lambda_i\|$, $|\partial_{\pi^2}l_{it}(\pi_{it})|$ and $|\partial_{\pi^3}l_{it}(\pi_{it})|$ are all bounded within $B(\mathcal{D})$, we have

$$
\begin{aligned}
|L1i| &\le T\left\|\hat{\lambda}_i-\lambda_i^G\right\|^2 M,\\
|L2i| &\le T^{\frac{1}{2}}\left\|\hat{\lambda}_i-\lambda_i^G\right\|\left\|\hat{f}-f^G\right\| M,\\
|L3i| &\le \left\|\hat{f}-f^G\right\|^2 M,
\end{aligned}
$$

for some large M w.p.a.1. Thus by Proposition 4 and Theorem 1, $|L1i|$, $|L2i|$ and $|L3i|$ are all $O_p(\frac{T}{\delta^2_{NT}})$, and $\sum_{i=1}^N|L1i|$, $\sum_{i=1}^N|L2i|$ and $\sum_{i=1}^N|L3i|$ are all $O_p(\frac{NT}{\delta^2_{NT}})$.

Now consider $(\hat{\phi}-\phi^G)'\partial_{\phi\phi'\lambda_{iq}}P(\phi^*_{iq})(\hat{\phi}-\phi^G)$. From equations (20), (21) and (22), it can be verified that

$$
\begin{aligned}
\partial_{\phi\phi'\lambda_{iq}}\Big[\Big(\frac{\sum_{i=1}^N\lambda_{iq}^2}{N}-\frac{\sum_{t=1}^T f_{tq}^2}{T}\Big)^2\Big] &= 8D_{NT}^{-1}(1_{iq}v_q'+v_q1_{iq}')D_{NT}^{-1}\\
&\quad +\frac{8}{N}\lambda_{iq}D_{NT}^{-1}(I_{N+T}\otimes\iota_q),\\
\partial_{\phi\phi'\lambda_{iq}}\Big[\sum\nolimits_{p=1}^r\sum\nolimits_{q=p+1}^r\Big(\sum\nolimits_{t=1}^T f_{tp}f_{tq}\Big)^2\Big] &= 0,\\
\partial_{\phi\phi'\lambda_{iq}}\Big[\sum\nolimits_{p=1}^r\sum\nolimits_{q=p+1}^r\Big(\sum\nolimits_{i=1}^N \lambda_{ip}\lambda_{iq}\Big)^2\Big] &= 2\Big(\sum\nolimits_{p\neq q}\lambda_{ip}D_1+\sum\nolimits_{p\neq q}1_{ip}u_{pq}'\Big)\\
&\quad +2\sum\nolimits_{p\neq q}u_{pq}1_{ip}'.
\end{aligned}
$$

1_{iq} is an $Nr + Tr$ dimensional vector with the q-th element in the i-th block being one and all the other elements being zero. Thus $(\hat{\phi} - \phi^G)' \partial_{\phi\phi'\lambda_{iq}} P(\phi^*_{iq})(\hat{\phi} - \phi^G)$ equals

$$
\begin{aligned}
& cNT(\hat{\phi} - \phi^G)' D_{NT}^{-1}(1_{iq} v_q' + v_q 1_{iq}') D_{NT}^{-1}(\hat{\phi} - \phi^G) \\
& + cT(\hat{\phi} - \phi^G)' \lambda_{iq} D_{NT}^{-1}(I_{N+T} \otimes \iota_q)(\hat{\phi} - \phi^G) \\
& + \frac{cT}{N}(\hat{\phi} - \phi^G)' \sum\nolimits_{p \neq q} \lambda_{ip} D_1(\hat{\phi} - \phi^G) + \frac{cT}{N}(\hat{\phi} - \phi^G)' \sum\nolimits_{p \neq q} (1_{ip} u_{pq}' + u_{pq} 1_{ip}')(\hat{\phi} - \phi^G) \\
\equiv \; & P1i + P2i + P3i + P4i.
\end{aligned}
$$

It follows that

$$
\begin{aligned}
|P1i| &= 2cNT \left| \frac{1}{N}(\hat{\lambda}_{iq} - \lambda_{iq}^G)[\frac{1}{N} \sum\nolimits_{j=1}^N \lambda_{jq}(\hat{\lambda}_{jq} - \lambda_{jq}^G) - \frac{1}{T} \sum\nolimits_{t=1}^T f_{tq}(\hat{f}_{tq} - f_{tq}^G)] \right| \\
&\leq MT \left\| \hat{\lambda}_i - \lambda_i^G \right\| (\frac{1}{N} \|\lambda\| \left\| \hat{\lambda} - \lambda^G \right\| + \frac{1}{T} \|f\| \left\| \hat{f} - f^G \right\|), \\
|P2i| &\leq MT \|\lambda_i\| (\frac{1}{N} \left\| \hat{\lambda} - \lambda^G \right\|^2 + \frac{1}{T} \left\| \hat{f} - f^G \right\|^2), \\
|P3i| &\leq \frac{cT}{N} \left\| \sum\nolimits_{p \neq q} \lambda_{ip} \iota_{pq} \right\| \left\| \hat{\lambda} - \lambda^G \right\|^2 \leq M \frac{T}{N} \|\lambda_i\| \left\| \hat{\lambda} - \lambda^G \right\|^2, \\
|P4i| &= \frac{2cT}{N} \left| \sum\nolimits_{p \neq q} (\hat{\lambda}_{ip} - \lambda_{ip}^G)(\sum\nolimits_{j=1}^N \lambda_{jq}(\hat{\lambda}_{jp} - \lambda_{jp}^G) + \sum\nolimits_{j=1}^N \lambda_{jp}(\hat{\lambda}_{jq} - \lambda_{jq}^G)) \right| \\
&\leq M \frac{T}{N} \left\| \hat{\lambda}_i - \lambda_i^G \right\| \|\lambda\| \left\| \hat{\lambda} - \lambda^G \right\|.
\end{aligned}
$$

Thus by Proposition 4, Theorem 1, $|P1i|$, ..., $|P4i|$ are all $O_p(\frac{T}{\delta_{NT}^2})$, while $\sum_{i=1}^N |P1i|$, ..., $\sum_{i=1}^N |P4i|$ are all $O_p(\frac{NT}{\delta_{NT}^2})$.

Taking together, we have shown $\|[R_\lambda]_i\|_1 = O_p(\frac{T}{\delta_{NT}^2})$ and $\|R_\lambda\|_1 = O_p(\frac{NT}{\delta_{NT}^2})$. The other half of the Lemma follows from symmetry. ■

Lemma 10 *Under Assumptions 1-4 and 6-8, as* $(N, T) \to \infty$,

(i) $(\hat{F} - F^G)' z = O_p(\frac{T}{\delta_{NT}^2} N^{\frac{3}{\xi}} T^{\frac{4}{\xi}})$,

(ii) $(\hat{F} - F^G)' \epsilon = O_p(\frac{T}{\delta_{NT}^2} N^{\frac{3}{\xi}} T^{\frac{4}{\xi}})$.

Proof. Part (i): From equation (8), we have $\hat{f}_t - f_t^G = [\hat{\phi} - \phi^G]_{N+t} = -[H^{-1}S]_{N+t} - \frac{1}{2}[H^{-1}R]_{N+t}$. It follows that

$$
-(\hat{F} - F^G)' z = \sum\nolimits_{t=1}^T [H^{-1}S]_{N+t} z_t' + \frac{1}{2} \sum\nolimits_{t=1}^T [H^{-1}R]_{N+t} z_t'.
$$

First consider the second term on the right hand side. The (q,j)-th element is $\sum_{t=1}^{T}[H^{-1}R]_{(N+t)q}z_{tj}$ and its magnitude is bounded by $\|[H^{-1}R]_f\|_1 \max_{t,j}|z_{tj}|$, where $[H^{-1}R]_f$ is the vector that contains the last Tr elements of $H^{-1}R$. From Assumption 8, it's easy to see that $\max_{t,j}|z_{tj}|$ is $O_p(T^{\frac{1}{\xi}})$. By Lemma 5 and Lemma 9, $\|[H^{-1}R]_f\|_1$ is $O_p(\frac{T}{\delta_{NT}^2}N^{\frac{3}{\xi}}T^{\frac{3}{\xi}})$. Thus the second term on the right hand side is $O_p(\frac{T}{\delta_{NT}^2}N^{\frac{3}{\xi}}T^{\frac{4}{\xi}})$.

The first term on the right hand side is $O_p(N^{-\frac{1}{2}}T^{\frac{1}{2}})$. From equation (57) and by symmetry, the first term equals

$$\begin{aligned}&\sum_{t=1}^{T}[H_{ff'}^{-1}S_f]_t z_t' + \sum_{t=1}^{T}[H_{ff'}^{-1}H_{f\lambda'}(H_{\lambda\lambda'}-H_{\lambda f'}H_{ff'}^{-1}H_{f\lambda'})^{-1}H_{\lambda f'}H_{ff'}^{-1}S_f]_t z_t'\\&-\sum_{t=1}^{T}[H_{ff'}^{-1}H_{f\lambda'}H_{\lambda\lambda'}^{-1}S_\lambda]_t z_t'\\&-\sum_{t=1}^{T}[H_{ff'}^{-1}H_{f\lambda'}(H_{\lambda\lambda'}-H_{\lambda f'}H_{ff'}^{-1}H_{f\lambda'})^{-1}H_{\lambda f'}H_{ff'}^{-1}H_{f\lambda'}H_{\lambda\lambda'}^{-1}S_\lambda]_t z_t'. \qquad (47)\end{aligned}$$

Similar to expression (43), $\left\|(H_{\lambda\lambda'}-H_{\lambda f'}H_{ff'}^{-1}H_{f\lambda'})^{-1}\right\|$ is $O_p(T^{-1})$. From Assumption 8, it's easy to see that $\|z\| = O_p(T^{\frac{1}{2}})$. These together with parts (xi) and (xvi) of Lemma 4 and parts (ii) and (v) of Lemma 8 implies that the second to the fourth terms of expression (47) are all $O_p(N^{-\frac{1}{2}}T^{\frac{1}{2}})$.

From equation (38) and by symmetry, the first term of expression (47) equals

$$\begin{aligned}&\sum_{t=1}^{T}[H_{Lff'}^{-1}S_f]_t z_t' -\\&\sum_{t=1}^{T}[H_{Lff'}^{-1}(V_f,U_f)[-\frac{T}{cN}I_{r^2}+(V_f,U_f)'H_{Lff'}^{-1}(V_f,U_f)]^{-1}(V_f,U_f)'H_{Lff'}^{-1}S_f]_t z_t' \qquad (48)\end{aligned}$$

Similar to parts (iii) and (vii) of Lemma 4, $\|(V_f,U_f)\|$ is $O(T^{\frac{1}{2}})$ and $\left\|H_{Lff'}^{-1}\right\|$ is $O_p(N^{-1})$. Similar to expression (39), $\left\|[-\frac{T}{cN}I_{r^2}+(V_f,U_f)'H_{Lff'}^{-1}(V_f,U_f)]^{-1}\right\|$ is $O(NT^{-1})$. By Assumption 8, $\|z\|_F$ is $O_p(T^{\frac{1}{2}})$. These together with Lemma 8(i) implies that the second term of expression (48) is $O_p(N^{-\frac{1}{2}}T^{\frac{1}{2}})$.

Now consider the first term of (48). Its (q,j)-th element is $\sum_{t=1}^{T}z_{tj}[H_{Lff'}^{-1}S_f]_{tq}$, which equals $\sum_{t=1}^{T}z_{tj}1_q^{(r)\prime}(\sum_{i=1}^{N}\partial_{\pi^2}l_{it}\lambda_i^G\lambda_i^{G\prime})^{-1}(\sum_{i=1}^{N}\partial_\pi l_{it}\lambda_i^G)$. In Lemma 8(i), we have shown (by symmetry) that $\sum_{t=1}^{T}f_{tp}^G 1_q^{(r)\prime}(\sum_{i=1}^{N}\partial_{\pi^2}l_{it}\lambda_i^G\lambda_i^{G\prime})^{-1}(\sum_{i=1}^{N}\partial_\pi l_{it}\lambda_i^G)$ is $O_p(N^{-\frac{1}{2}}T^{\frac{1}{2}})$. Lemma 8(i) uses Assumption 5(ii). Here from Assumption 8(iii) we have $\sum_{i=1}^{N}\sum_{t=1}^{T}(\sum_{i=1}^{N}\partial_{\pi^2}l_{it}\lambda_i^G\lambda_i^{G\prime})^{-1}\lambda_i^G z_t'\partial_\pi l_{it} = O_p(N^{-\frac{1}{2}}T^{\frac{1}{2}})$. Thus the first term of expression (48) is $O_p(N^{-\frac{1}{2}}T^{\frac{1}{2}})$.

Part (ii): The proof is similar to part (i), with z_t replaced by ϵ_{t+h}. ■

Lemma 11 *Under Assumptions 1-4 and 6-7,* $(\hat{F}-F^G)'F = O_p(\frac{T}{\delta_{NT}^2}N^{\frac{3}{\xi}}T^{\frac{3}{\xi}})$ *as* $(N,T) \to \infty$*.*

Proof. The proof is the same as the proof of Lemma 10, with z replaced by F^G. The result still holds if we replace F by $\partial_\pi l_{i\cdot}$, where $\partial_\pi l_{i\cdot} = (\partial_\pi l_{i1}, ..., \partial_\pi l_{iT})'$. ■

C Proof of Propositions and Theorems

C.1 Proof of Proposition 1

Proof. The essence of the following proof is that a low rank matrix can not fit a high rank matrix, which also underlies the proof of consistency in Bai (2009). The key technique (using the expansion of $L(X \left| \hat{f}, \hat{\lambda}\right.)$ and boundedness from below of $-\partial_{\pi^2} l_{it}(\cdot)$) is inspired by Chen et al. (2014).

Expand $l_{it}(\hat{\pi}_{it})$ at π_{it}^0, we have $l_{it}(\hat{\pi}_{it}) = l_{it}(\pi_{it}^0) + \partial_\pi l_{it} \times (\hat{\pi}_{it} - \pi_{it}^0) + \frac{1}{2}\partial_{\pi^2} l_{it}(\pi_{it}^*) \times (\hat{\pi}_{it} - \pi_{it}^0)^2$. It follows that

$$
\begin{aligned}
L(X \left| \hat{f}, \hat{\lambda}\right.) &= \sum\nolimits_{i=1}^{N} \sum\nolimits_{t=1}^{T} l_{it}(\hat{\pi}_{it}) = \sum\nolimits_{i=1}^{N} \sum\nolimits_{t=1}^{T} l_{it}(\pi_{it}^0) \\
&\quad + \sum\nolimits_{i=1}^{N} \sum\nolimits_{t=1}^{T} \partial_\pi l_{it} \times (\hat{\pi}_{it} - \pi_{it}^0) \qquad (49) \\
&\quad + \frac{1}{2} \sum\nolimits_{i=1}^{N} \sum\nolimits_{t=1}^{T} \partial_{\pi^2} l_{it}(\pi_{it}^*) \times (\hat{\pi}_{it} - \pi_{it}^0)^2. \qquad (50)
\end{aligned}
$$

$|\pi_{it}^*|$ is bounded because by design $|\hat{\pi}_{it}|$ is bounded and by Assumption 1 $|\pi_{it}^0|$ is also bounded. Thus $|\partial_{\pi^2} l_{it}(\pi_{it}^*)|$ is bounded below by b_L. Expression (50) is negative and its absolute value is not smaller than $\frac{b_L}{2} \sum_{i=1}^{N} \sum_{t=1}^{T} (\hat{\pi}_{it} - \pi_{it}^0)^2$. From inequality $|tr(AB')| \leq rank(B) \left\| A \right\| \left\| B \right\|_F$, the absolute value of expression (49) is not larger than $2r \left\| \partial_\pi l \right\| [\sum_{i=1}^{N} \sum_{t=1}^{T} (\hat{\pi}_{it} - \pi_{it}^0)^2]^{\frac{1}{2}}$, where $\partial_\pi l$ is $N \times T$ matrix with $\partial_\pi l_{it}$ in the i-th row and t-th column.

Next, since $P(\hat{f}, \hat{\lambda}) \leq 0$ and $P(f^G, \lambda^G) = 0$, we have $L(X \left| \hat{f}, \hat{\lambda}\right.) \geq Q(\hat{f}, \hat{\lambda})$ and $L(X \left| f^G, \lambda^G\right.) = Q(f^G, \lambda^G)$. By definition, $Q(\hat{f}, \hat{\lambda}) \geq Q(f^G, \lambda^G)$, thus $L(X \left| \hat{f}, \hat{\lambda}\right.) \geq$

$L(X\,|f^G,\lambda^G)$. It follows that expression (49) must be positive and not smaller than the absolute value of expression (50). Thus from Lemma 1 we have

$$[\sum_{i=1}^{N}\sum_{t=1}^{T}(\hat{\pi}_{it}-\pi_{it}^0)^2]^{\frac{1}{2}} \le \frac{4r}{b_L}\left\|\partial_\pi l\right\| = O_p(N^{\frac{1}{2}}T^{\frac{1}{4}}+N^{\frac{1}{4}}T^{\frac{1}{2}}). \tag{51}$$

Recall that $\rho_1,...,\rho_r$ and $\hat{\rho}_1,...,\hat{\rho}_r$ are singular values of $N^{-\frac{1}{2}}T^{-\frac{1}{2}}F^G\Lambda^{G\prime}$ and $N^{-\frac{1}{2}}T^{-\frac{1}{2}}\hat{F}\hat{\Lambda}'$ respectively. Let $e_1,...,e_r$ and $\hat{e}_1,...,\hat{e}_r$ be the corresponding left-singular vectors. From Davis-Kahan Theorem (see Stewart and Sun (1990)), for $j=1,...,r$ we have,

$$\left\|\hat{e}_j-e_j\right\| \le \sqrt{2}\left\|N^{-\frac{1}{2}}T^{-\frac{1}{2}}\hat{F}\hat{\Lambda}'-N^{-\frac{1}{2}}T^{-\frac{1}{2}}F^G\Lambda^{G\prime}\right\|/\eta, \tag{52}$$

where $\eta=\min\{\left|\rho_{j-1}-\hat{\rho}_j\right|\wedge\left|\rho_{j+1}-\hat{\rho}_j\right|, j=1,...,r\}$. From equation (51) we have $\left\|N^{-\frac{1}{2}}T^{-\frac{1}{2}}\hat{F}\hat{\Lambda}'-N^{-\frac{1}{2}}T^{-\frac{1}{2}}F^G\Lambda^{G\prime}\right\| \le N^{-\frac{1}{2}}T^{-\frac{1}{2}}\left\|\hat{F}\hat{\Lambda}'-F^G\Lambda^{G\prime}\right\|_F = O_p(N^{-\frac{1}{4}}+T^{-\frac{1}{4}})$. Note that $\rho_1,...,\rho_r$ are all bounded and bounded away from zero in probability. Thus by Weyl's inequality, $|\rho_1-\hat{\rho}_1|,...,|\rho_r-\hat{\rho}_r|$ are all $O_p(N^{-\frac{1}{4}}+T^{-\frac{1}{4}})$. Thus from Assumptions 6 we can conclude η is bounded and bounded away from zero in probability. It follows that (52) implies $\left\|\hat{e}_j-e_j\right\|=O_p(N^{-\frac{1}{4}}+T^{-\frac{1}{4}})$. For $j=1...,r$, under penalty function (3) the j-th estimated factor is $\sqrt{T\hat{\rho}_j}\hat{e}_j$. Under condition (4), the j-th factor is $\sqrt{T\rho_j}e_j$. Thus we have

$$\left\|\hat{f}-f^G\right\| \le \sqrt{T}\left|\sqrt{\hat{\rho}_j}-\sqrt{\rho_j}\right|\left\|\hat{e}_j\right\|+\sqrt{T\rho_j}\left\|\hat{e}_j-e_j\right\| = T^{\frac{1}{2}}O_p(N^{-\frac{1}{4}}+T^{-\frac{1}{4}}).$$

By symmetry, $\left\|\hat{\lambda}-\lambda^G\right\| = N^{\frac{1}{2}}O_p(N^{-\frac{1}{4}}+T^{-\frac{1}{4}})$. ■

C.2 Proof of Proposition 2 and Proposition 3

Proof. Let $\hat{\phi}^*$ be the solution of the problem: $\min_{\phi\in B(\mathcal{D})\cap\left\|D_{NT}^{-\frac{1}{2}}(\phi-\phi^G)\right\|\le m}\left\|D_{TN}^{-\frac{1}{2}}S(\phi)\right\|_\zeta$. ζ is defined in Assumption 4. Expand $S(\hat{\phi}^*)$ at ϕ^G using integral form of the mean value theorem, we have

$$\hat{\phi}^*-\phi^G=\tilde{H}^{-1}(S(\hat{\phi}^*)-S). \tag{53}$$

(1): By Lemma 7, the infinity norm of the upper-left block of $\tilde{H}^{-1}$ (which equals

1-norm since $\tilde{H}^{-1}$ is symmetric) is $O_p(\frac{N^{\frac{2}{\xi}}T^{\frac{2}{\xi}}}{T})$, and the infinity norm of the upper-right block of $\tilde{H}^{-1}$ (which equals 1-norm of the lower-left block of $\tilde{H}^{-1}$) is $O_p(\frac{N^{\frac{3}{\xi}}T^{\frac{3}{\xi}}}{N})$. $\left\|T^{-\frac{1}{2}}(S_\lambda(\hat{\phi}^*) - S_\lambda)\right\|_\infty$ is $O_p(N^{\frac{1}{\zeta}} + T^{\frac{1}{\zeta}})$ because

$$
\begin{aligned}
\left\|T^{-\frac{1}{2}}(S_\lambda(\hat{\phi}^*) - S_\lambda)\right\|_\infty &\leq \left\|T^{-\frac{1}{2}}(S_\lambda(\hat{\phi}^*) - S_\lambda)\right\|_\zeta \\
&\leq \left\|D_{TN}^{-\frac{1}{2}}(S(\hat{\phi}^*) - S)\right\|_\zeta \leq 2\left\|D_{TN}^{-\frac{1}{2}}S\right\|_\zeta = O_p((N+T)^{\frac{1}{\zeta}}),
\end{aligned}
$$

where the last equality follows from Assumption 4. Similarly, $\left\|N^{-\frac{1}{2}}(S_f(\hat{\phi}^*) - S_f)\right\|_\infty$ is also $O_p((N+T)^{\frac{1}{\zeta}})$. Thus by Assumption 7,

$$
\begin{aligned}
\left\|\hat{\lambda}^* - \lambda^G\right\|_\infty &= T^{-\frac{1}{2}}O_p(N^{\frac{2}{\xi}}T^{\frac{2}{\xi}}(N+T)^{\frac{1}{\zeta}}) = o_p(1), \\
\left\|\hat{f}^* - f^G\right\|_\infty &= N^{-\frac{1}{2}}O_p(N^{\frac{3}{\xi}}T^{\frac{3}{\xi}}(N+T)^{\frac{1}{\zeta}}) = o_p(1).
\end{aligned}
$$

(2): By equation (53), $D_{NT}^{-\frac{1}{2}}(\hat{\phi}^* - \phi^G) = N^{-\frac{1}{2}}T^{-\frac{1}{2}}(D_{TN}^{-\frac{1}{2}}\tilde{H}D_{TN}^{-\frac{1}{2}})^{-1}[D_{TN}^{-\frac{1}{2}}(S(\hat{\phi}^*) - S)]$. By Holder's inequality,

$$
\left\|D_{TN}^{-\frac{1}{2}}(S(\hat{\phi}^*) - S)\right\| \leq (N+T)^{\frac{1}{2}-\frac{1}{\zeta}}\left\|D_{TN}^{-\frac{1}{2}}(S(\hat{\phi}^*) - S)\right\|_\zeta = O_p((N+T)^{\frac{1}{2}}).
$$

This together with Lemma 3 shows that $\left\|D_{NT}^{-\frac{1}{2}}(\hat{\phi}^* - \phi^G)\right\| = O_p(N^{-\frac{1}{2}} + T^{-\frac{1}{2}})$.

(3): Part (1) implies that $\hat{\phi}^*$ is an interior point of $B(\mathcal{D})$ w.p.a.1, because ϕ^G lies in $B(\frac{\mathcal{D}}{2})$ w.p.a.1. Part (2) implies that $\hat{\phi}^*$ is an interior point of $\left\|D_{NT}^{-\frac{1}{2}}(\phi - \phi^G)\right\| \leq m$ w.p.a.1. Thus $\hat{\phi}^*$ is an interior point of $B(\mathcal{D})\cap\left\|D_{NT}^{-\frac{1}{2}}(\phi - \phi^G)\right\| \leq m$ w.p.a.1. By definition of $\hat{\phi}^*$, this implies that $\partial_\phi\left\|D_{TN}^{-\frac{1}{2}}S(\phi)\right\|_\zeta^\zeta\Big|_{\phi=\hat{\phi}^*} = 0$.

It follows that $[D_{TN}^{-\frac{1}{2}}S(\hat{\phi}^*)]^{\zeta-1} = 0$ since $\partial_\phi\left\|D_{TN}^{-\frac{1}{2}}S(\phi)\right\|_\zeta^\zeta = \zeta H(\phi)D_{TN}^{-\frac{1}{2}}[D_{TN}^{-\frac{1}{2}}S(\phi)]^{\zeta-1}$ (here $[D_{TN}^{-\frac{1}{2}}S(\phi)]^{\zeta-1}$ denotes the vector that each element equals the $\zeta - 1$ power of the corresponding element of $D_{TN}^{-\frac{1}{2}}S(\phi)$), and by Lemma 3, $D_{TN}^{-\frac{1}{2}}H(\phi)D_{TN}^{-\frac{1}{2}}$ is negative definite within $B(\mathcal{D})\cap\left\|D_{NT}^{-\frac{1}{2}}(\phi - \phi^G)\right\| \leq m$ w.p.a.1.

Thus we also have $S(\hat{\phi}^*) = 0$ w.p.a.1, and consequently $\hat{\phi}^*$ is the unique maximizer of the likelihood within $B(\mathcal{D})\cap\left\|D_{NT}^{-\frac{1}{2}}(\phi - \phi^G)\right\| \leq m$ w.p.a.1.

(4): By definition, $\hat{\phi}$ maximizes the likelihood within $B(\mathcal{D})$. Proposition 1 shows that $\hat{\phi}$ lies in the neighborhood $\left\| D_{NT}^{-\frac{1}{2}}(\phi - \phi^G) \right\| \leq m$ w.p.a.1. Thus $\hat{\phi}$ maximizes the likelihood within $B(\mathcal{D}) \cap \left\| D_{NT}^{-\frac{1}{2}}(\phi - \phi^G) \right\| \leq m$ w.p.a.1.

(5): Part (3) and part (4) together implies that $\hat{\phi} = \hat{\phi}^*$ w.p.a.1. In the following, we simply use $\hat{\phi}$ to denote both. ■

C.3 Proof of Theorem 1

Proof. As explained in the main context, Theorem 1 follows from equation (5), Lemma 3 and $\left\| D_{TN}^{-\frac{1}{2}} S \right\| = O_p((N+T)^{\frac{1}{2}})$. ■

C.4 Proof of Proposition 4

Proof. The first term on the right hand side of equation (6) equals

$$\begin{aligned} & \sum\nolimits_{t=1}^{T} (\partial_\pi l_{it}) f_t^G + \sum\nolimits_{t=1}^{T} [\partial_\pi l_{it}(\hat{f}_t' \lambda_i^G) - \partial_\pi l_{it}] f_t^G \\ & + \sum\nolimits_{t=1}^{T} (\partial_\pi l_{it})(\hat{f}_t - f_t^G) + \sum\nolimits_{t=1}^{T} [\partial_\pi l_{it}(\hat{f}_t' \lambda_i^G) - \partial_\pi l_{it}](\hat{f}_t - f_t^G) \\ = \; & O_p(T^{\frac{1}{2}} + \frac{T}{\delta_{NT}}). \end{aligned}$$

By Assumption 5, the first term is $O_p(T^{\frac{1}{2}})$. By Assumption 2(ii), $(\sum_{t=1}^{T} [\partial_\pi l_{it}(\hat{f}_t' \lambda_i^G) - \partial_\pi l_{it}]^2)^{\frac{1}{2}}$ is not larger than $b_U \left\| \hat{f} - f^G \right\| \left\| \lambda_i^G \right\|$. By Theorem 1, $\left\| \hat{f} - f^G \right\|$ is $O_p(\frac{T^{\frac{1}{2}}}{\delta_{NT}})$. By Assumption 1, $\left\| f^G \right\|$ is $O_p(T^{\frac{1}{2}})$. Thus the second term is $O_p(\frac{T}{\delta_{NT}})$. $(\sum_{t=1}^{T} (\partial_\pi l_{it})^2)^{\frac{1}{2}}$ is $O_p(T^{\frac{1}{2}})$ because by Assumption 3(i), $\mathbb{E}(\partial_\pi l_{it})^2$ is uniformly bounded. Thus the third term is $O_p(\frac{T}{\delta_{NT}})$. The fourth term is $O_p(\frac{T}{\delta_{NT}^2})$ because it is not larger than $b_U \left\| \lambda_i^G \right\| \left\| \hat{f} - f^G \right\|^2$.

For the second term on the right hand side of equation (6), we have:

$$\begin{aligned} & \rho_{\min}(T^{-1} \sum\nolimits_{t=1}^{T} [\int_0^1 -\partial_{\pi^2} l_{it}(\hat{f}_t'(\lambda_i^G + s(\hat{\lambda}_i - \lambda_i^G))) ds] \hat{f}_t \hat{f}_t') \\ \geq \; & b_L \rho_{\min}(T^{-1} \sum\nolimits_{t=1}^{T} \hat{f}_t \hat{f}_t') \xrightarrow{p} b_L (\rho_{\min}(\Sigma_F \Sigma_\Lambda))^{\frac{1}{2}} > 0. \end{aligned}$$

The first inequality follows from Assumption 2(ii). "$\xrightarrow{p}$" follows from Theorem 1 and Assumption 1. Thus w.p.a.1, the norm of second term on the right hand side of equation (6) is not smaller than $\frac{1}{2}Tb_L(\rho_{\min}(\Sigma_F\Sigma_\Lambda))^{\frac{1}{2}}\left\|\hat{\lambda}_i-\lambda_i^G\right\|$. Thus $\left\|\hat{\lambda}_i-\lambda_i^G\right\|$ is $T^{-1}O_p(T^{\frac{1}{2}}+\frac{T}{\delta_{NT}})$, which is $O_p(\frac{1}{\delta_{NT}})$. By symmetry, $\left\|\hat{f}_t-f_t^G\right\|$ is also $O_p(\frac{1}{\delta_{NT}})$. ■

C.5 Proof of Theorem 2

Proof. As presented in the main text, $\hat{\lambda}_i-\lambda_i^G=[\hat{\phi}-\phi^G]_i=-[H^{-1}S]_i-\frac{1}{2}[H^{-1}R]_i$. First consider $[H^{-1}R]_i$. From equation (40) we have

$$\begin{aligned}&[H^{-1}R]_i=[H_{\lambda\lambda'}^{-1}R_\lambda]_i+[H_{\lambda\lambda'}^{-1}H_{\lambda f'}(H_{ff'}-H_{f\lambda'}H_{\lambda\lambda'}^{-1}H_{\lambda f'})^{-1}H_{f\lambda'}H_{\lambda\lambda'}^{-1}R_\lambda]_i\\&-[H_{\lambda\lambda'}^{-1}H_{\lambda f'}H_{ff'}^{-1}R_f]_i\\&-[H_{\lambda\lambda'}^{-1}H_{\lambda f'}(H_{ff'}-H_{f\lambda'}H_{\lambda\lambda'}^{-1}H_{\lambda f'})^{-1}H_{f\lambda'}H_{\lambda\lambda'}^{-1}H_{\lambda f'}H_{ff'}^{-1}R_f]_i.\end{aligned}\tag{54}$$

Consider the four terms one by one.

(R1): From equation (38) we have

$$\begin{aligned}[H_{\lambda\lambda'}^{-1}R_\lambda]_i &= [H_{L\lambda\lambda'}^{-1}]_i[R_\lambda]_i-[H_{L\lambda\lambda'}^{-1}]_i[(V_\lambda,U_\lambda)]_i[-\frac{N}{cT}I_{r^2}+\\&\quad (V_\lambda,U_\lambda)'H_{L\lambda\lambda'}^{-1}(V_\lambda,U_\lambda)]^{-1}(V_\lambda,U_\lambda)'H_{L\lambda\lambda'}^{-1}R_\lambda\end{aligned}\tag{55}$$

By Lemma 4(v), and Lemma 9, the 1-norm of the first term on the right hand side of equation (55) is $O_p(\frac{1}{\delta_{NT}^2})$. By inequality (39), Lemma 9 and parts (i), (ii), (v) and (vi) of Lemma 4, the 1-norm of the second term on the right hand side of equation (55) is also $O_p(\frac{1}{\delta_{NT}^2})$. Taking together, we have $\left\|[H_{\lambda\lambda'}^{-1}R_\lambda]_i\right\|_1=O_p(\frac{1}{\delta_{NT}^2})$.

(R2): From equation (38), the second term on the right hand side of (54) equals

$$\begin{aligned}&[H_{L\lambda\lambda'}^{-1}]_i[H_{\lambda f'}(H_{ff'}-H_{f\lambda'}H_{\lambda\lambda'}^{-1}H_{\lambda f'})^{-1}H_{f\lambda'}]_iH_{\lambda\lambda'}^{-1}R_\lambda\\&-[H_{L\lambda\lambda'}^{-1}]_i[(V_\lambda,U_\lambda)]_i[-\frac{N}{cT}I_{r^2}+(V_\lambda,U_\lambda)'H_{L\lambda\lambda'}^{-1}(V_\lambda,U_\lambda)]^{-1}(V_\lambda,U_\lambda)'H_{L\lambda\lambda'}^{-1}\\&\times H_{\lambda f'}(H_{ff'}-H_{f\lambda'}H_{\lambda\lambda'}^{-1}H_{\lambda f'})^{-1}H_{f\lambda'}H_{\lambda\lambda'}^{-1}R_\lambda.\end{aligned}\tag{56}$$

By equation (42), $\left\|[H_{\lambda f'}(H_{ff'}-H_{f\lambda'}H_{\lambda\lambda'}^{-1}H_{\lambda f'})^{-1}H_{f\lambda'}]_i\right\|_1$ is $O_p(N^{\frac{2}{\xi}-1}T^{\frac{2}{\xi}+1})$. This together with parts (v) and (viii) of Lemma 4 and Lemma 9 implies that the 1-norm of the first term of expression (56) is $O_p(\frac{N^{\frac{2}{\xi}}T^{\frac{2}{\xi}}}{\delta_{NT}^2})$. For the second term of expression (56), parts (i), (ii), (v), (vi) and (viii) of Lemma 4, equation (41), inequality (39) and Lemma 9 together implies that the 1-norm of this term is also $O_p(\frac{N^{\frac{2}{\xi}}T^{\frac{2}{\xi}}}{\delta_{NT}^2})$. Taking together, the 1-norm of expression (56) is $O_p(\frac{N^{\frac{2}{\xi}}T^{\frac{2}{\xi}}}{\delta_{NT}^2})$.

(R3): The 1-norm of the third term on the right hand side of (54) is $\frac{N^{\frac{1}{\xi}}T^{\frac{1}{\xi}}}{\delta_{NT}^2}$. The calculation procedure is similar to (R1). The difference is that R_λ is replaced by $H_{\lambda f'}H_{ff'}^{-1}R_f$. Part (R1) uses $\|[R_\lambda]_i\|_1 = O_p(\frac{T}{\delta_{NT}^2})$ and $\|R_\lambda\|_1 = O_p(\frac{NT}{\delta_{NT}^2})$. Here by Lemma 9 and parts (x), (xiv) and (xv) of Lemma 4, $\left\|[H_{\lambda f'}H_{ff'}^{-1}R_f]_i\right\|_1$ is $O_p(\frac{T^{\frac{1}{\xi}+1}}{\delta_{NT}^2})$ and $\left\|H_{\lambda f'}H_{ff'}^{-1}R_f\right\|_1$ is $O_p(\frac{N^{\frac{1}{\xi}+1}T^{\frac{1}{\xi}+1}}{\delta_{NT}^2})$.

(R4): The 1-norm of the fourth term on the right hand side of (54) is $O_p(\frac{N^{\frac{3}{\xi}}T^{\frac{3}{\xi}}}{\delta_{NT}^2})$. The calculation procedure is similar to (R2). The difference is R_λ is replaced by $H_{\lambda f'}H_{ff'}^{-1}R_f$.

Taking (R1)-(R4) together, we have $\|[H^{-1}R]_i\|_1 = O_p(\frac{N^{\frac{3}{\xi}}T^{\frac{3}{\xi}}}{\delta_{NT}^2})$. Since $[H^{-1}R]_i$ is a fixed dimensional vector, its 1-norm and Euclidean norm has the same order, thus $\|[H^{-1}R]_i\|$ is also $O_p(\frac{N^{\frac{3}{\xi}}T^{\frac{3}{\xi}}}{\delta_{NT}^2})$. Note that here we choose to calculate $\|[H^{-1}R]_i\|_1$ rather than $\|[H^{-1}R]_i\|$ directly, because calculating $\|[H^{-1}R]_i\|$ requires calculating $\|[R]_i\|$ and $\|R\|$. From term "$L1i$" in Lemma 9, we can see that this requires calculating the exact rate of $\left\|\hat{\lambda} - \lambda^G\right\|_4$, which seems quite difficult and tedious.

Now consider $[H^{-1}S]_i$. From equation (40) we have

$$
\begin{aligned}
&[H^{-1}S]_i = [H_{\lambda\lambda'}^{-1}S_\lambda]_i + [H_{\lambda\lambda'}^{-1}H_{\lambda f'}(H_{ff'} - H_{f\lambda'}H_{\lambda\lambda'}^{-1}H_{\lambda f'})^{-1}H_{f\lambda'}H_{\lambda\lambda'}^{-1}S_\lambda]_i \\
&- [H_{\lambda\lambda'}^{-1}H_{\lambda f'}H_{ff'}^{-1}S_f]_i \\
&-[H_{\lambda\lambda'}^{-1}H_{\lambda f'}(H_{ff'} - H_{f\lambda'}H_{\lambda\lambda'}^{-1}H_{\lambda f'})^{-1}H_{f\lambda'}H_{\lambda\lambda'}^{-1}H_{\lambda f'}H_{ff'}^{-1}S_f]_i. \qquad (57)
\end{aligned}
$$

Consider the four terms one by one.

(S1): From equation (38) we have

$$
\begin{aligned}
[H_{\lambda\lambda'}^{-1}S_\lambda]_i &= [H_{L\lambda\lambda'}^{-1}S_\lambda]_i - [H_{L\lambda\lambda'}^{-1}]_i[(V_\lambda, U_\lambda)]_i[-\frac{N}{cT}I_{r^2} + \\
&\quad (V_\lambda, U_\lambda)'H_{L\lambda\lambda'}^{-1}(V_\lambda, U_\lambda)]^{-1}(V_\lambda, U_\lambda)'H_{L\lambda\lambda'}^{-1}S_\lambda \qquad (58)
\end{aligned}
$$

Consider the second term on the right hand side of equation (58). Since $[H^{-1}_{L\lambda\lambda'}]_i$ is symmetric, we have $\left\|[H^{-1}_{L\lambda\lambda'}]_i\right\| \leq (\left\|[H^{-1}_{L\lambda\lambda'}]_i\right\|_1 \left\|[H^{-1}_{L\lambda\lambda'}]_i\right\|_\infty)^{\frac{1}{2}} = \left\|[H^{-1}_{L\lambda\lambda'}]_i\right\|_1$. Thus by Lemma 4(v), $\left\|[H^{-1}_{L\lambda\lambda'}]_i\right\|$ is $O_p(T^{-1})$. $\|[V_\lambda, U_\lambda]_i\|$ is $O_p(1)$. These together with equation (39) and Lemma 8(i) implies that the norm of the second term is $O_p(N^{-\frac{1}{2}}T^{-\frac{1}{2}})$. Thus we have $[H^{-1}_{\lambda\lambda'}S_\lambda]_i = [H^{-1}_{L\lambda\lambda'}S_\lambda]_i + O_p(N^{-\frac{1}{2}}T^{-\frac{1}{2}})$.

(S2): From equation (38), the second term on the right hand side of (57) equals

$$
\begin{aligned}
&[H^{-1}_{L\lambda\lambda'}]_i[H_{\lambda f'}]_i(H_{ff'} - H_{f\lambda'}H^{-1}_{\lambda\lambda'}H_{\lambda f'})^{-1}H_{f\lambda'}H^{-1}_{\lambda\lambda'}S_\lambda \\
&-[H^{-1}_{L\lambda\lambda'}]_i[(V_\lambda, U_\lambda)]_i[-\frac{N}{cT}I_{r^2} + (V_\lambda, U_\lambda)'H^{-1}_{L\lambda\lambda'}(V_\lambda, U_\lambda)]^{-1}(V_\lambda, U_\lambda)'H^{-1}_{L\lambda\lambda'} \\
&\times H_{\lambda f'}(H_{ff'} - H_{f\lambda'}H^{-1}_{\lambda\lambda'}H_{\lambda f'})^{-1}H_{f\lambda'}H^{-1}_{\lambda\lambda'}S_\lambda. \qquad (59)
\end{aligned}
$$

As explained in (S1), $\left\|[H^{-1}_{L\lambda\lambda'}]_i\right\|$ is $O_p(T^{-1})$. This together with Lemma 4(xiii), inequality (43) and parts (ii) and (v) of Lemma 8 implies that the norm of the first term of (59) is $O_p(N^{-\frac{1}{2}}T^{-\frac{1}{2}})$.

Next consider the second term of (59). As explained in (S1), $\left\|[H^{-1}_{L\lambda\lambda'}]_i\right\|$ is $O_p(T^{-1})$ and $\|[V_\lambda, U_\lambda]_i\|$ is $O_p(1)$. These together with equation (39), parts (iii), (vii) and (xvi) of Lemma 4, inequality (43) and parts (ii) and (v) of Lemma 8 implies that the norm of the second term of (59) is $O_p(N^{-\frac{1}{2}}T^{-\frac{1}{2}})$. Taking together, the norm of expression (59) is $O_p(N^{-\frac{1}{2}}T^{-\frac{1}{2}})$.

(S3): The norm of the third term on the right hand side of (57) is $O_p(N^{-\frac{3}{4}}T^{-\frac{1}{4}})$.

$$
\begin{aligned}
&[H^{-1}_{\lambda\lambda'}H_{\lambda f'}H^{-1}_{ff'}S_f]_i = [H^{-1}_{L\lambda\lambda'}]_i[H_{\lambda f'}H^{-1}_{ff'}S_f]_i - [H^{-1}_{L\lambda\lambda'}]_i[(V_\lambda, U_\lambda)]_i[-\frac{N}{cT}I_{r^2} + \\
&(V_\lambda, U_\lambda)'H^{-1}_{L\lambda\lambda'}(V_\lambda, U_\lambda)]^{-1}(V_\lambda, U_\lambda)'H^{-1}_{L\lambda\lambda'}H_{\lambda f'}H^{-1}_{ff'}S_f. \qquad (60)
\end{aligned}
$$

As explained in (S1), $\left\|[H^{-1}_{L\lambda\lambda'}]_i\right\|$ is $O_p(T^{-1})$. This together with parts (iii) and (iv) of Lemma 8 implies the norm of the first term on the right hand side of (60) is $O_p(N^{-\frac{1}{2}}T^{-\frac{1}{2}})$.

The norm of the second term on the right hand side of (60) is also $O_p(N^{-\frac{1}{2}}T^{-\frac{1}{2}})$. The calculation procedure is similar to the second term on the right hand side of equation (58). The difference is that S_λ is replaced by $H_{\lambda f'}H^{-1}_{ff'}S_f$. (S1) uses $\left\|(V_\lambda, U_\lambda)'H^{-1}_{L\lambda\lambda'}S_\lambda\right\| = O_p(N^{\frac{1}{2}}T^{-\frac{1}{2}})$. Here due to parts (iii) and (vii) of Lemma 4 and parts (ii) and (v) of Lemma 8, $\left\|(V_\lambda, U_\lambda)' H_{L\lambda\lambda'}^{-1} H_{\lambda f'} H_{ff'}^{-1} S_f\right\|$ is also $O_p(N^{\frac{1}{2}}T^{-\frac{1}{2}})$.

(S4): The norm of the fourth term on the right hand side of (57) is $O_p(N^{-\frac{1}{2}}T^{-\frac{1}{2}})$. The calculation procedure is similar to (S2). The difference is that S_λ is replaced by $H_{\lambda f'} H_{ff'}^{-1} S_f$. (S2) uses $\left\|H_{f\lambda'} H_{\lambda\lambda'}^{-1} S_\lambda\right\| = O_p(N^{\frac{1}{2}})$. Here due to parts (ix) and (xvi) of Lemma 4 and parts (ii) and (v) of Lemma 8, $\left\|H_{f\lambda'} H_{\lambda\lambda'}^{-1} H_{\lambda f'} H_{ff'}^{-1} S_f\right\|$ is also $O_p(N^{\frac{1}{2}})$.

Taking (S1)-(S4) together, we have $[H^{-1}S]_i = [H_{L\lambda\lambda'}^{-1} S_\lambda]_i + O_p(N^{-\frac{1}{2}}T^{-\frac{1}{2}})$. Thus

$$\hat{\lambda}_i - \lambda_i^G = -[H_{L\lambda\lambda'}^{-1} S_\lambda]_i + O_p(N^{-\frac{1}{2}}T^{-\frac{1}{2}}) + O_p(\frac{N^{\frac{3}{\xi}}T^{\frac{3}{\xi}}}{\delta_{NT}^2}).$$

By Assumption 5(iii), we have $T[H_{L\lambda\lambda'}^{-1}]_i \xrightarrow{p} (\bar{G}'\Sigma_{iF}\bar{G})^{-1}$ and $T^{-\frac{1}{2}}[S_\lambda]_i \xrightarrow{d} \mathcal{N}(0, \bar{G}'\Omega_{iF}\bar{G})$. Since $\frac{T^{\frac{1}{2}}}{\delta_{NT}^2} N^{\frac{3}{\xi}} T^{\frac{3}{\xi}} \to 0$, we have $T^{\frac{1}{2}}(\hat{\lambda}_i - \lambda_i^G) \xrightarrow{d} \mathcal{N}(0, \bar{G}^{-1}\Sigma_{iF}^{-1}\Omega_{iF}\Sigma_{iF}^{-1}\bar{G}'^{-1})$. Limit distribution of estimated factors follows from symmetry. Consistency of var_λ and var_f follows from Assumption 2(ii), Assumption 3 and Theorem 1. ■

C.6 Proof of Proposition 5

Proof. For expression (12): The j-th diagonal element of $T^{-1}\hat{F}'\hat{F}$ and $T^{-1}F^{G\prime}F^G$ is $\hat{\rho}_j$ and ρ_j respectively.

$$\begin{aligned}
\left\|\mathcal{V}_{NT}^{\frac{1}{2}} - \mathcal{V}^{\frac{1}{2}}\right\| &= \max_j |\hat{\rho}_j - \rho_j| \leq (\sum\nolimits_{j=1}^r (\hat{\rho}_j - \rho_j)^2)^{\frac{1}{2}} \leq T^{-1}\left\|\hat{F}'\hat{F} - F^{G\prime}F^G\right\|_F \\
&\leq 2\left\|(\hat{F} - F^G)'F^G\right\|_F + \left\|\hat{F} - F^G\right\|_F^2
\end{aligned}$$

Thus by Lemma 11 and Theorem 1, $\left\|\mathcal{V}_{NT}^{\frac{1}{2}} - \mathcal{V}^{\frac{1}{2}}\right\|$ is $O_p(\frac{N^{\frac{3}{\xi}}T^{\frac{3}{\xi}}}{\delta_{NT}^2})$. Since $\hat{\rho}_j$ and ρ_j are all bounded and bounded away from zero in probability, $\|\mathcal{V}_{NT} - \mathcal{V}\|$, $\left\|\mathcal{V}_{NT}^{\frac{1}{4}} - \mathcal{V}^{\frac{1}{4}}\right\|$ and $\left\|\mathcal{V}_{NT}^{-1} - \mathcal{V}^{-1}\right\|$ are all $O_p(\frac{N^{\frac{3}{\xi}}T^{\frac{3}{\xi}}}{\delta_{NT}^2})$.

For expression (13): First note that $G\mathcal{V}_{NT}^{-\frac{1}{4}} = (\frac{\Lambda^{0\prime}\Lambda^0}{N})^{\frac{1}{2}}\Upsilon\mathcal{V}^{-\frac{1}{4}}\mathcal{V}_{NT}^{-\frac{1}{4}}$ and $H_{Bai} \equiv \frac{\Lambda^{0\prime}\Lambda^0}{N}\frac{F^{0\prime}\tilde{F}}{T}\mathcal{V}_{NT}^{-1}$. It suffices to show $\left\|\frac{F^{0\prime}\tilde{F}}{T} - (\frac{\Lambda^{0\prime}\Lambda^0}{N})^{-\frac{1}{2}}\Upsilon\mathcal{V}^{\frac{1}{2}}\right\| = O_p(\frac{N^{\frac{3}{\xi}}T^{\frac{3}{\xi}}}{\delta_{NT}^2})$. Noting that $\frac{F^{0\prime}F^0G\mathcal{V}^{-\frac{1}{4}}}{T} = (\frac{\Lambda^{0\prime}\Lambda^0}{N})^{-\frac{1}{2}}\Upsilon\mathcal{V}^{\frac{1}{2}}$, it suffices to show $\left\|\frac{1}{T}F^{0\prime}(\tilde{F} - F^0G\mathcal{V}^{-\frac{1}{4}})\right\| = O_p(\frac{N^{\frac{3}{\xi}}T^{\frac{3}{\xi}}}{\delta_{NT}^2})$. This can be proved by the following facts: (1) $\mathcal{V}_{NT}^{-\frac{1}{4}} - \mathcal{V}^{-\frac{1}{4}} = -\mathcal{V}_{NT}^{-\frac{1}{4}}(\mathcal{V}_{NT}^{\frac{1}{4}} - \mathcal{V}^{\frac{1}{4}})\mathcal{V}^{-\frac{1}{4}}$, (2) $\tilde{F} - F^0G\mathcal{V}^{-\frac{1}{4}} = (\hat{F} - F^0G)\mathcal{V}_{NT}^{-\frac{1}{4}} + F^0G(\mathcal{V}_{NT}^{-\frac{1}{4}} - \mathcal{V}^{-\frac{1}{4}})$, (3) by Lemma 11, $\frac{1}{T}F^{0\prime}(\hat{F} - F^G) =$

$O_p(\frac{N^{\frac{3}{\xi}}T^{\frac{3}{\xi}}}{\delta_{NT}^2})$. ■

C.7 Proof of Theorem 3

Proof. Let $Y = (y_{1+h}, ..., y_{T+h})'$, $z = (z_1, ..., z_T)'$, $\hat{z} = (\hat{z}_1, ..., \hat{z}_T)'$, $\epsilon = (\epsilon_{1+h}, ..., \epsilon_{T+h})'$. It follows that $Y = \hat{z}\delta + \epsilon + (F^G - \hat{F})G^{-1}\alpha$ and $\hat{\delta} = (\hat{z}'\hat{z})^{-1}\hat{z}'Y = \delta + (\hat{z}'\hat{z})^{-1}(\hat{z}'\epsilon + \hat{z}'(F^G - \hat{F})G^{-1}\alpha)$. Let $\Xi = diag(G, I_q)$, then we have $\hat{z} - z\Xi = \hat{F} - F^G$. Then due to facts listed below, we have $\hat{\delta} - \delta = (\Xi'z'z\Xi + O_p(\frac{T}{\delta_{NT}}))^{-1}(\Xi'z'\epsilon + O_p(\frac{T}{\delta_{NT}^2}N^{\frac{3}{\xi}}T^{\frac{4}{\xi}}))$. By Assumption 8, $T^{-1}z'z \xrightarrow{p} \Sigma_{zz}$ and $T^{-\frac{1}{2}}z'\epsilon \xrightarrow{d} \mathcal{N}(0, \Sigma_{zz\epsilon})$. Thus given $\frac{T^{\frac{1}{2}}}{\delta_{NT}^2}N^{\frac{3}{\xi}}T^{\frac{4}{\xi}} \to 0$, we have $T^{\frac{1}{2}}(\hat{\delta} - \delta) \xrightarrow{d} \mathcal{N}(0, \bar{\Xi}^{-1}\Sigma_{zz}^{-1}\Sigma_{zz\epsilon}\Sigma_{zz}^{-1}\bar{\Xi}'^{-1})$. The proof for consistency of $\hat{\Sigma}_\delta$ is straightforward and hence omitted.

(1) $\epsilon'\hat{z} = \epsilon'z\Xi + \epsilon'(\hat{F} - F^G)$.

(2) $\hat{z}'(F^G - \hat{F}) = (\hat{F} - F^G)'(F^G - \hat{F}) + \Xi'z'(F^G - \hat{F})$.

(3) By Lemma 10, both $(\hat{F} - F^G)'z$ and $(\hat{F} - F^G)'\epsilon$ are $O_p(\frac{T}{\delta_{NT}^2}N^{\frac{3}{\xi}}T^{\frac{4}{\xi}})$.

(4) By Theorem 1, $\hat{F} - F^G$ is $O_p(\frac{T^{\frac{1}{2}}}{\delta_{NT}})$.

(5) Both $\|\Xi\|$ and $\|G^{-1}\|$ are $O_p(1)$.

(6) By Assumption 8(i), $\|z\|$ is $O_p(T^{\frac{1}{2}})$. ■

C.8 Proof of Theorem 4

Proof. First, $\hat{y}_{T+h|T} - y_{T+h|T} = z_T'\Xi(\hat{\delta} - \delta) + (\hat{f}_T - f_T^G)'G^{-1}\alpha + (\hat{z}_T - \Xi'z_T)'(\hat{\delta} - \delta)$.

By Theorem 3, $T^{\frac{1}{2}}z_T'\Xi(\hat{\delta} - \delta) \xrightarrow{d} \mathcal{N}(0, z_T'\Sigma_{zz}^{-1}\Sigma_{zz\epsilon}\Sigma_{zz}^{-1}z_T)$.

By Theorem 2, under the assumption $\frac{N^{\frac{1}{2}}}{\delta_{NT}^2}N^{\frac{3}{\xi}}T^{\frac{3}{\xi}} \to 0$, we have $N^{\frac{1}{2}}(\hat{f}_T - f_T^G)'G^{-1}\alpha \xrightarrow{d} \mathcal{N}(0, \alpha'\Sigma_{t\Lambda}^{-1}\Omega_{t\Lambda}\Sigma_{t\Lambda}^{-1}\alpha)$.

$(\hat{z}_T - \Xi'z_T)'(\hat{\delta} - \delta)$ is $O_p(\frac{1}{T^{\frac{1}{2}}\delta_{NT}})$ because (1) by Theorem 4, $\|\hat{z}_T - \Xi'z_T\| = \left\|\hat{f}_T - f_T^G\right\| = O_p(\frac{1}{\delta_{NT}})$; (2) under the assumption $\frac{T^{\frac{1}{2}}}{\delta_{NT}^2}N^{\frac{3}{\xi}}T^{\frac{4}{\xi}} \to 0$, Theorem 3 shows that $\left\|\hat{\delta} - \delta\right\|$ is $O_p(T^{-\frac{1}{2}})$.

By Assumption 8, ϵ_t is independent with x_{is} for all i and s, thus $z_T'\Xi(\hat{\delta} - \delta)$ is asymptotically uncorrelated with $(\hat{f}_T - f_T^G)'G^{-1}\alpha$.

These together implies that $(\hat{y}_{T+h|T} - y_{T+h|T})/B_T \xrightarrow{d} \mathcal{N}(0, 1)$. The proof for consistency of $\hat{B}_T^2$ is straightforward. ■

D Verification of Assumption 2

(1) Probit:

The likelihood function is $l_{it}(\pi) = x_{it}\log\Phi(\pi) + (1-x_{it})\log(1-\Phi(\pi))$. The first order derivative is $\partial_\pi l_{it}(\pi) = x_{it}\frac{\phi(\pi)}{\Phi(\pi)} - (1-x_{it})\frac{\phi(\pi)}{1-\Phi(\pi)}$. The second order derivative is $\partial_{\pi^2} l_{it}(\pi) = x_{it}(\frac{-\pi\phi(\pi)}{\Phi(\pi)} - \frac{\phi^2(\pi)}{\Phi^2(\pi)}) - (1-x_{it})(\frac{-\pi\phi(\pi)}{1-\Phi(\pi)} + \frac{\phi^2(\pi)}{(1-\Phi(\pi))^2})$. Let $m(\pi) = \frac{\phi(\pi)}{\Phi(\pi)}$ be the inverse mill's ratio and $q_{it} = 2x_{it} - 1$. It follows that $l_{it}(\pi) = \log\Phi(q_{it}\pi)$, $\partial_\pi l_{it}(\pi) = q_{it}m(q_{it}\pi)$ and $\partial_{\pi^2} l_{it}(\pi) = -q_{it}\pi m(q_{it}\pi) - m^2(q_{it}\pi)$. Now consider a standard normal random variable truncated on the right at $q_{it}\pi$. Its variance is $1 - q_{it}\pi m(q_{it}\pi) - m^2(q_{it}\pi)$. Since $|q_{it}\pi| = |\pi|$ is bounded, the variance must be strictly greater than zero and less than one. Thus $-\partial_{\pi^2} l_{it}(\pi)$ is also strictly greater than zero and less than one. The third order derivative is $\partial_{\pi^3} l_{it}(\pi) = -q_{it}[m(q_{it}\pi) + q_{it}\pi m'(q_{it}\pi) + 2m(q_{it}\pi)m'(q_{it}\pi)]$. Since $|q_{it}\pi| = |\pi|$ is bounded, $|m(q_{it}\pi)|$ and $|m'(q_{it}\pi)|$ are also bounded. Thus $|\partial_{\pi^3} l_{it}(\pi)|$ is also bounded.

(2) Logit:

The likelihood function is $l_{it}(\pi) = x_{it}\log\Psi(\pi) + (1-x_{it})\log(1-\Psi(\pi))$, where $\Psi(\pi) = \frac{e^\pi}{1+e^\pi}$. The first order derivative is $\partial_\pi l_{it}(\pi) = x_{it}\frac{\Psi'(\pi)}{\Psi(\pi)} - (1-x_{it})\frac{\Psi'(\pi)}{1-\Psi(\pi)}$, which equals $x_{it} - \Psi(\pi)$ once we plugging in $\Psi'(\pi) = \Psi(\pi) - \Psi^2(\pi)$. The second order derivative is $\partial_{\pi^2} l_{it}(\pi) = -\Psi(\pi)(1-\Psi(\pi))$. The third order derivative is $\partial_{\pi^3} l_{it}(\pi) = \Psi(\pi)(2\Psi(\pi)-1)(1-\Psi(\pi))$. It is easy to see that given $|\pi|$ is bounded, $-\partial_{\pi^2} l_{it}(\pi)$ is less than or equal to $\frac{1}{4}$ and strictly greater than zero, and $|\partial_{\pi^3} l_{it}(\pi)|$ is less than 1.

(3) Poisson:

The likelihood function[11] is $l_{it}(\pi) = -e^\pi + k\pi - \log k!$ because $P(x_{it} = k) = p(k,\lambda) = e^{-\lambda}\lambda^k/k!$ where $\lambda = e^\pi$. The first order derivative is $\partial_\pi l_{it}(\pi) = -e^\pi + k$. Both the second and third order derivatives are $\partial_{\pi^2} l_{it}(\pi) = -e^\pi$. Thus it is easy to see that given $|\pi|$ is bounded, $-\partial_{\pi^2} l_{it}(\pi)$ and $|\partial_{\pi^3} l_{it}(\pi)|$ are both bounded away from zero and bounded above.

(4) Tobit:

Since Tobit represents a class of models, we show through a representative case.

[11] Note that for Poisson regression, $\lambda = e^\pi$ rather than $\lambda = \pi$. λ has to be positive while π could be negative. In standard Poisson regression, $\pi = x'\theta$, while here $\pi_{it} = f_t^{0\prime}\lambda_i^0$.

Suppose $x_{it}^* = \pi_{it} + e_{it}$, e_{it} is $N(0,1)$, and $x_{it} = x_{it}^*$ if $x_{it}^* > 0$, $x_{it} = 0$ if $x_{it}^* \leq 0$. The likelihood function is $l_{it}(\pi) = -\frac{1}{2}(x_{it} - \pi)^2 \mathbf{1}(x_{it} > 0) + \log(1 - \Phi(\pi))\mathbf{1}(x_{it} = 0)$, where $\mathbf{1}(\cdot)$ is the indicator function. The second order derivative is $\partial_{\pi^2} l_{it}(\pi) = -1$ if $x_{it} > 0$, $\partial_{\pi^2} l_{it}(\pi) = -(-\pi m(-\pi) + m^2(-\pi))$ if $x_{it} = 0$. The third order derivative is $\partial_{\pi^3} l_{it}(\pi) = 0$ if $x_{it} > 0$, $\partial_{\pi^3} l_{it}(\pi) = m(-\pi) - \pi m'(-\pi) + 2m(-\pi)m'(-\pi)$ if $x_{it} = 0$. These together with the argument in the Probit case shows $-\partial_{\pi^2} l_{it}(\pi)$ is bounded away from zero and both $-\partial_{\pi^2} l_{it}(\pi)$ and $|\partial_{\pi^3} l_{it}(\pi)|$ are bounded above.